R842-05

ERIC REPORT RESUME

ED 010 258   2-28-67   24   (REV)

TRANSLATED READING TESTS AS CULTURE-FAIR MEASURES FOR FOREIGN STUDENTS.

KUMBARACI, TURKAN E.
DCW10512 COLUMBIA UNIV., NEW YORK, N.Y.

CRP-S-177
BR-5-8214
 -   -66   OEC-5-10-108
EDRS PRICE MF-$0.18 HC-$4.96   124P.

CULTURAL DIFFERENCES, *READING COMPREHENSION, *READING TESTS, *SCREENING TESTS, *FOREIGN STUDENTS, TURKISH, COLLEGE ENTRANCE EXAMINATIONS, COLLEGE STUDENTS, *TEST VALIDITY, COMPARATIVE ANALYSIS, *ITEM ANALYSIS, *CULTURE FREE TESTS, TURKEY, NEW YORK CITY, NEW YORK

A COMPARISON OF AN ENGLISH LANGUAGE READING COMPREHENSION TEST WITH ITS TURKISH TRANSLATION AND RETRANSLATION WAS CONDUCTED. THE INSTRUMENTS CONSISTED OF TWO PARALLEL FORMS OF A READING TEST OF COLLEGE ENTRANCE LEVEL. THEY WERE TRANSLATED INTO TURKISH, AND THEN RETRANSLATED INTO ENGLISH. SUPPLEMENTARY MEASURES WERE ALSO EMPLOYED. THE SAMPLE CONSISTED OF 896 TURKISH HIGH SCHOOL SENIORS AND COLLEGE STUDENTS, AND 1,324 AMERICAN HIGH SCHOOL SENIORS AND COLLEGE STUDENTS. SEVERAL SUGGESTIONS WERE DISCUSSED FOR THE PERFECTION OF THE INSTRUMENTS USED FOR SCREENING FOREIGN STUDENTS AND FOR CROSS-CULTURAL ITEM STATISTICS. (RS)

5-8214
(5-177)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

TRANSLATED READING TESTS AS CULTURE-FAIR MEASURES FOR FOREIGN STUDENTS

by

Turkan Emine Kumbaraci

ED010258

This report which is being submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Columbia University was performed pursuant to Contract number OE-5-10-108 with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program.

1966

TABLE OF CONTENTS

Chapter Page

I--PROBLEM AND BACKGROUND 1

The Concept of Culture-Fairness in International Evaluations 4

Shortcomings of Previous Research in International Evaluations of the Culture-Fairness of Aptitude Tests 7

Emphasis upon Performance Tests for Practical Details of Administration 7

Considering Verbal and Performance Tests Interchangeable Aptitude Measures 8

Confounding Foreign Language and Cultural Effects in Verbal Tests 9

Confounding Verbal Test Adaptation with Cultural Effects 10

Methods Used in International Evaluations of Culture-Fairness 12

Comparisons Based on Total Scores 12

Selection of Items by Use of Empirical Data from Different Countries 13

Selection of Content from Different National Sources and Comparison of Item Characteristics 14

Relationship of Methods Used to Present Study 16

Chapter Page

Shortcomings of English Proficiency Measures as a Solution to the Language-Fairness Problem 17
Low Predictive Validity for General Achievement . . . 18
Difference in Processes Involved in Various Levels of Foreign Language Learning and Fluency in the Vernacular 19
Implications for Present Study 22
Problem Investigated by Present Study 22
Theoretical Query: Culture-Fairness 22
Practical Query: Language-Fairness 24
Related Studies 24
Assumptions 25
II--DESIGN AND PROCEDURE 28
Instruments 28
International Reading Test 28
Practice Test for Turkish Students Tested in the Vernacular 31
Vocabulary Test for Turkish Students Studying English 31
Questionnaire for Turkish Students Studying English. . 31
Supplementary Measures 32
SAT-Verbal Scores 32
Intelligence Test Scores 32
University of Ankara Entrance Examination Scores 32
Grades in Literature 33
First Semester Grades in an American University . 33

Chapter Page

Sample 34

  Turkish High School Seniors 34

  Turkish College Students 39

  American High School Seniors 39

  American College Students 42

  Turkish Students Studying English 44

Administration Procedure 45

  Turkey 46

  United States 47

  Turkish Students Studying English 47

Analyses of Data 48

  Scoring 48

  Analysis of Variance 48

  Subsample Analysis 49

  Correlation of Item Difficulty Indices 49

  Correlation of Item Discrimination Indices 51

  Correlation of Difficulty and Discrimination Indices . 51

  Correlation of Popularity of Errors 52

  Evaluation of the Relative Across-Country Difficulty of Specific Items and Item Responses 52

  Relative Difficulty of Reading Passages 54

  Indices of Reliability 54

    Internal Consistency 55

Chapter Page

Stability of Difficulty and Discrimination Indices within One Country 55

Alternate-Form Reliability 55

Alternate-Language Reliability 55

Correlation with Other Measures 55

III--RESULTS AND CONCLUSIONS 57

Total Test Scores 57

Item Difficulty 62

Item Discrimination 66

Popularity of Errors 70

Specific Items and Item Responses 72

Reading Passages 81

Reliability 81

Correlation with Other Measures 85

IV--DISCUSSION 90

V--SUMMARY 101

References 104

Appendix 110

LIST OF TABLES

Table | Page

1 Analysis of Total Population in School and Increase in Enrollment between 1953-1957 in the United States and Turkey 35

2 Turkish High School Sample with Respect to Rank on the Criterion, School Type, and Students Tested 38

3 Means and Standard Deviations of Intelligence Scores for American High School Seniors Taking the Original English Versions 40

4 Means and Standard Deviations of Intelligence Scores for American High School Seniors Taking the Original and Re-Translated English Versions 40

5 Means and Standard Deviations of Intelligence Scores for American High School Seniors in Academic and General Curricula 41

6 Means and Standard Deviations of SAT-Verbal Scores for American College Students 43

7 Age Distribution of American College Students 44

8 Means and Standard Deviations of Total Scores for High School and College Students in the United States and Turkey . . . 58

9 Analysis of the Means and Standard Deviations of Total Scores for American High School Students by Curriculum 58

10 Analysis of Variance of Total Scores on the Two Forms Obtained by American and Turkish College Students 59

11 Analysis of Variance of Total Scores on the Two Forms Obtained by American and Turkish High School Students . . . 59

12 Means and Standard Deviations of Total Scores for the Original and Re-Translated English Versions Administered to American Students 60

13 Analysis of Variance of Total Scores on the Original and Re-Translated English Versions of the two Parallel Forms . . . 60

Table Page

14 Means and Standard Deviations of Total Scores for Turkish Students Studying English on the Reading Test and the Vocabulary Test G-T 61

15 Analysis of Variance of Total Scores on the Two Turkish Forms with Differential Testing Order 63

16 Correlations of Item Difficulty Indices for American College and Turkish High School Subsamples 64

17 Correlations of Item Difficulty Indices for Original and Re-Translated English Versions Administered to American Students 66

18 Correlations of Item Difficulty Indices for English Versions Administered to American College Students and Turkish Students Studying English 67

19 Correlations of Item Discrimination Indices for American College and Turkish High School Subsamples 68

20 Within-Country Correlations between Difficulty and Discrimination Indices 69

21 Correlations of Errors for American College and Turkish High School Subsamples 71

22 Correlations of Errors for Original and Re-Translated English Versions Administered to American and Turkish Samples 71

23 Relationship between Relative Item Difficulty and Discrimination for American College and Turkish High School Samples 73

24 Examples of Items with No Overall Difficulty or Error Difference 75

25 Examples of Items with No Overall Difficulty Difference Containing Different Popular Errors 76

26 Examples of Items with Overall Difficulty Difference Containing Different Popular Errors 77

27 Examples of Items with Overall Difficulty Difference also Containing Different Popular Errors 78

28 Correlation of Difficulty of Reading Passages for American College and Turkish High School Samples 82

Table Page

29 Internal Consistency of the Original and Re-Translated English and Translated Turkish Versions 83

30 Alternate-Form and Alternate-Language Reliabilities for Turkish Samples 85

31 Correlation of Scores on the Original and Re-Translated English Versions with SAT-Verbal Scores and IQ 86

32 Correlation of Scores on the Translated Turkish Versions with Grades in Literature 87

CHAPTER I

PROBLEM AND BACKGROUND

In recent years there has been a rapid increase in the number of foreign students continuing their education in the United States. The 82,000 foreign students in 1965 represents an increase of 900 per cent since 1930. Indeed, there was an increase of 10 per cent between 1964 and 1965.[1] Approximately half these students are college undergraduates; a third are graduate students. At present, 6 per cent of all graduate students and about 2 per cent of all undergraduates in the United States are from foreign countries.[2]

This expansion in international student exchange is paralleled by an increasing necessity for, and interest in, language-and culture-fair appraisals of foreign student aptitude. The issues of language- and culture-fairness are interrelated, but pose quite distinct problems.

The language-fairness issue arises from the fact that American screening devices such as the College Entrance Examination Board Tests and the Graduate Record Examinations are not in the vernacular of foreign students. Scores on especially the verbal subtests do not reveal the differential effects of language proficiency and academic aptitude, since students vary in the length of English training and competence in its use.

---

[1] Institute of International Education, Open doors, 1965: Report on international exchange. New York: Author, 1965. Pp. 4-6.

[2] Committee on Educational Interchange Policy, College and university programs of academic exchange--suggestions for the study of exchanges of students, faculty and short-term visitors. New York: Author, 1960. P. 7.

The SAT verbal score in English has very little predictive validity for foreign student academic achievement, as stated in a report of workshops sponsored by the College Entrance Examination Board and the Institute of International Education.[1] This may be attributed to the contamination of scores by the level of English competence. Therefore, admission officers either give more weight to the SAT quantitative score as a predictor of foreign student achievement or omit the SAT verbal score completely.

Furthermore, English competence is currently considered a prerequisite for higher education in the United States, but not a variable influencing the selection process for foreign students. Many American universities and organizations have facilities for, and are willing to undertake a program of language instruction for foreign students after a screening process. According to Harris,[2] about 20 per cent of American colleges which have fewer than 30 foreign students, and about 10 per cent of those which have more than 30 foreign students do not require any evidence of English proficiency before students leave their own countries. It has also been mentioned that the command of English required of these students may vary depending upon their major fields of study. For example, a foreign student studying sociology needs to know more English than another studying biology; the latter needs to be more competent in English than another studying sculpture.[1] Due to all these factors, a recent conference sponsored by the Center for Applied Linguistics[2] recognized the necessity not only for a central testing program for the growing number of foreign students, but also for the construction of aptitude tests for students with a minimal knowledge of English or zero English proficiency.

---

[1] The College Entrance Examination Board and the Institute of International Education, U.S. college and university policies, practices, and problems in admitting foreign students. New York: Institute of International Education, 1965. Pp. 18-34.

[2] D. P. Harris, A survey of English language requirements and facilities for foreign students in United States institutions of higher learning, 1961. New York: National Association for Foreign Student Affairs, 1962.

While the problem of language-fairness may be solved merely by translating the tests into the vernacular of foreign students, determination of the culture-fairness of a test depends upon stringent evaluations of test content. Content may be defined either in terms of specific questions asked, or in terms of test format. The content of a culture-fair test presents equivalent stimulus materials to students from different cultures and, in turn, draws equivalent responses from different cultural groups.

The present study explores the culture-fairness of translated versions of a college level reading test in order to determine if further translations of this kind could be used as language and culture-fair devices in the screening and guidance of foreign students.

---

[1] Committee on the Foreign Student in American Colleges and Universities, _The college, the university and the foreign student_. New York: Author, 1963. P. 14.

[2] Center for Applied Linguistics of the Modern Language Association of America, _Testing the English proficiency of foreign students_. Washington, D. C.: Author, 1961. Pp. 2-4.

The Concept of Culture-Fairness in International Evaluations

Defined broadly, "culture is the learned portion of human behavior."[1] More specifically, "the culture of a people consists of their distinctive modal patterns of behavior and the underlying regulatory beliefs, values, norms, and premises."[2] Using this definition, Anastasi[3] remarked that no test can be culture-free, because all behavior is moulded by cultural factors and all tests measure aspects of behavior.

Yet, it is possible to construct a culture-fair test. According to Kretch, Crutchfield and Ballachey,[4] some behavioral patterns are universal; others belong distinctively to a particular group of people. Within this framework, a culture-fair test is one which is composed of elements shared by many cultures, and which omits those elements peculiar only to one culture.

Kretch, Crutchfield and Ballachey also make a distinction between the following two elements, which influence the design of studies dealing with culture-fairness:

1. The "particular set of cultural arrangements adopted by a society,"[4] where different countries constitute different societies.

---

[1] M. J. Herskovits, Cultural anthropology. New York: Alfred A. Knopf, 1960. P. 313.

[2] D. Kretch, R. S. Crutchfield, and E. L. Ballachey, Individual in society. New York: McGraw Hill, 1962. P. 344.

[3] Anne Anastasi, Psychological testing. New York: Macmillan, 1959. P. 255.

[4] D. Kretch, R. S. Crutchfield, & E. L. Ballachey, Individual in society. P. 341.

In this sense, a culture-fair test would be comprised of elements which are similar for comparable groups in different countries, but which distinguish between educational-socioeconomic strata in each of the respective countries.

2. The subcultures within each society, where "each social class carries and maintains a more or less distinct culture."[1] In this sense, a culture-fair test would not contrast different educational-socioeconomic strata in a particular country, but would contrast the performance of subjects from different countries.

This distinction is important because different elements seem to contrast social classes in a particular country and relatively similar social classes in different countries.

Verbal factors, such as reading vocabulary, reading comprehension, and knowledge of the standard language differentiate social classes in a particular country. Evidence can be found in Carroll,[2] Davis,[3] and Sexton.[4] Studies by Stevanovic[5] and Pieter[6] conducted in

---

[1] D. Kretch, R. S. Crutchfield, & E. L. Ballachey, Individual in society. P. 372.

[2] J. B. Carroll, The study of language. Cambridge, Mass: Harvard University Press, 1953.

[3] Allison Davis, Social class influences upon learning. Cambridge, Mass.: Harvard University Press, 1955.

[4] Patricia C. Sexton, Education and income. New York: Viking Press, 1961. P. 27.

[5] B. P. Stevanovic, The development of the child's intelligence and the Beograd revision of the Binet-Simon scale; summary data and results. Bull. Acad. Lettr. serbe, 1935, 1, 89-114.

[6] J. Pieter, Intelligence quotient and environment. Kwart. Psychol., 1939, 11, 265-322.

Europe using the Binet also support the above statement. This suggests that culture-fair tests for different educational-socioeconomic strata in a particular country would be those which include a minimum of verbal tasks.

On the other hand, it is not yet known which aspects of form and content make an important difference in changing the nature of a test for subjects from different countries. There have been only two suggestive conclusions:

1. The content of non-verbal or performance tests is not universally shared by all nations. Non-verbal tests are culture-fair only for educated subjects in different countries, because of the cultural proximity that develops through education. For example, Verhaegen[1] and Ombredane,[2] working in Africa with various performance tests such as the Kohs Cubes, Progressive Matrices, and Colored Matrices found the tasks involved to be culturally loaded for aborigines.

2. Subjects in high educational-socioeconomic strata in different countries obtain almost similar scores on verbal tests if the language handicap is reduced. McKillop and Yoloye[3] used the Vocabulary Test G-T with university students in Nigeria, and found that American and Nigerian college samples did equally well, even though English is

---

[1] F. Verhaegen, Utilite actuelle des tests pour l'etude psychologique des autochtones congolais. Rev. Psychol. appl., 1956, 6, 139-151.

[2] A. Ombredane, Etude du comportement intellectuel des noirs congolais. Psychol. franc., 1957, 1, 19.

[3] Anne McKillop, & E. A. Yoloye, The reading of university students. Teach. Educ., 1962, 3, 93-107.

not the vernacular in Nigeria.

These findings suggest that data from within-country studies on culture-fairness are not particularly relevant to across-country studies. Construction of international tests requires thorough exploration of elements shared by different countries.

Shortcomings of Previous Research in International Evaluations of the Culture-Fairness of Aptitude Tests

Several factors for which no allowances were made in previous research in cross-cultural testing prevent an evaluation of the results in terms of the culture-fairness of tests for different countries.

Emphasis upon Performance Tests for Practical Details of Administration

Emphasis upon language differences between countries has resulted in the construction of non-verbal or performance tests which can often be administered in pantomine. Examples of such tests are the International Group Mental Test by Dodd, the Leiter International Performance Scale, the Progressive Matrices Test by Raven, and the Navy-Northwestern Matrices Test.

It is true that non-verbal tests simplify administration procedures by eliminating the necessity for translation. It is also true that in cases where both non-verbal and verbal tests were administered in a foreign country, total scores on the former showed a better approximation to scores on the norming sample. Examples may be found in studies by Church,[1] and Garth, Elson, and Morton.[2]

Yet, practical facility does not imply inherent culture-fairness. Studies of this kind show merely the reduction of language handicap on non-verbal tests, since the verbal tests were administered in a foreign language. No inferences can be drawn from these studies about the relative culture-fairness of verbal type tests in contrast to performance type tests.

Considering Verbal and Performance Tests Interchangeable Aptitude Measures

There has been a tendency to contrast verbal and non-verbal aptitude tests in terms of culture-fairness. As shown by Choudhuri and Majumdar[3] in a factor-analytic study, verbal and performance tests cannot be considered interchangeable, because each measures different facets of intellect.

A test used for prediction is useful to the extent that it is a successful predictor. Non-verbal measures correlate lower with academic achievement than do verbal measures. This has been illustrated as early as 1922 by Gates[1] and recently confirmed by MacArthur and Elley,[2] Bennett, Seashore and Wesman,[3] and Lorge and Thorndike.[4] In a study by Bolton,[5] non-verbal I.Q.s predicted achievement only in courses such as mechanical drawing, printing, safety and health. Although they facilitated educational guidance when used with verbal I.Q.s, they increased the standard error of measurement. The similar results obtained by Keehn and Protho[6] in Lebanon suggest that these findings are not limited to the United States.

---

[1] A. M. Church, The standardized testing program summary report 1947. Hawaii educ. Rev., 1947, 36, 53-56.

[2] T. R. Garth, T. H. Elson, & M. M. Morton, The administration of non-language intelligence tests to Mexicans. J. abn. soc. Psychol., 1936, 31, 53-58.

[3] P. K. Choudhuri, & P. K. Majumdar, Factorial approach to the problem whether verbal intelligence tests can be replaced by performance type intelligence tests. Indian J. Psychol., 1963, 38, 125-128.

Confounding Foreign Language and Cultural Effects in Verbal Tests

In studies dealing with the administration of a test in a different country, using the test without translations or adaptations, the subjects are tested in a foreign language. Therefore, scores do not differentiate between effects due to language difficulty and cultural modes of response to test content. Examples of such studies are those of Westbrook,[1] Eaton,[2] and Coffman.[3] Coffman's analysis is interesting because item difficulties for American and African samples taking the tests were adjusted in order to identify the relatively easier and relatively more difficult items for each group. For example, African students found antonyms and analogies more difficult than items of the sentence completion type or reading comprehension. Coffman's analysis permits one to reach general conclusions dealing with test format. For a particular item where discrepant results were obtained from subjects in the two cultures, however, one cannot know if the change can be attributed to a difference in language competence or a difference between cultures.

---

[1] I. A. Gates, The correlation of achievement in school subjects with intelligence tests and other variables. J. educ. Psychol., 1922, 13, 129-139, 223-235, 277-285.

[2] R. S. MacArthur, & W. B. Elley, The reduction of socioeconomic bias in intelligence testing. Brit. J. educ. Psychol., 1963, 33, 107-119.

[3] G. K. Bennett, H. G. Seashore, & A. G. Wesman, Differential Aptitude Tests manual. New York: Psychological Corporation. 1959. P. 38.

[4] I. Lorge, & R. L. Thorndike, The Lorge-Thorndike Intelligence Tests technical manual. Boston: Houghton-Mifflin, 1962. P. 20.

[5] F. B. Bolton, Value of several intelligence tests for predicting scholastic achievement. J. educ. Res., 1947, 41, 133-138.

[6] J. D. Keehn, & E. T. Protho, Non-verbal tests as predictors of academic success in Lebanon. Educ. psychol. Measmt., 1955, 15, 495-498.

Confounding Verbal Test Adaptation with Cultural Effects

Adaptations such as modifying the wording of some items, eliminating other items, or changing the scoring methods have often been used along with the translation of a test into the language of a particular country in which it is going to be administered. There have been numerous studies in this area using the Stanford-Binet, the

---

[1] C. H. Westbrook, The use of English group intelligence tests with Chinese students. Education, Univ. Shanghai, 1940, 3, 83-95.

[2] M. T. Eaton, A survey of the language arts achievement of sixth grade children in 18 counties and 6 cities in India. Res. Bull. Ind. Dep. publ. Instruct., 1942, 3, 1-75.

[3] W. E. Coffman, Evidence of cultural factors in responses of African students to items in an American test of scholastic aptitude. Twentieth yearbook, National Council on Measurement in Education, 1963, 28-37.

Wechsler, and the Otis. The reader may refer to Kamat,[1] Malin,[2] Pasricha and Pagedar,[3] and Wu.[4] A recent report of workshops sponsored by the College Entrance Examination Board and the Institute of International Education[5] also mentioned the use of a Spanish version of the SAT in Puerto Rico where the test resembles the SAT in format, but very little in content.

A priori changes in test content in an effort to adapt the test for another culture transforms the original test to some degree. Thus it becomes difficult to compare scores on this modified test with scores on the original test in order to appraise its culture-fairness.

Of course, even in cases where only translations are used, a minimal amount of alteration in test content may be expected. According to Whorf, "we dissect nature along the lines laid down by our native languages. . . we cut nature up, organize it into concepts, and ascribe significances."[6] Thus an idea or a concept may not have a corresponding correlate in another language. Its difficulty cannot be predicted directly from another language.

---

[1] V. V. Kamat, A revision of the Binet scale for Indian children. Brit. J. educ. Psychol., 1934, 4, 296-309.

[2] A. J. Malin, An Indian adaptation of the WISC. J. voc. educ. Psychol., 1964, 10, 128-131.

[3] P. Pasricha, & R. M. Pagedar, Adaptation of "WAIS" to the Gujarati population. J. voc. educ. Guidance, 1963, 9, 174-184.

[4] T. M. Wu, On the second revision of the Chinese Binet-Simon scale. Shanghai: Commercial Press, 1936.

[5] The College Entrance Examination Board and the Institute of International Education, U.S. college and university policies, practices, and problems in admitting foreign students. New York: Institute of International Education, 1965. Pp. 18-34.

[6] B. Whorf, Science and linguistics. In H.B. Allen (Ed.), Readings in applied English linguistics. New York: Appleton-Century Crofts, 1958. P. 33.

However, Whorf also assumes that cognition and thought is moulded by the linguistic specifications of a particular culture. Cultural differences can at least partly be attributed to linguistic differences.

A test, adapted while translated, may from a practical standpoint become a valid measure for the country in question. Yet, data collected on this test do not distinguish between the variance attributed to cultural and linguistic differences as opposed to the variance due to a priori adaptations.

Methods Used in International Evaluations of Culture-Fairness

The determination of culture-fairness depends upon comparisons based on specific criteria. Various criteria have been used in international evaluations of this kind.

Comparisons Based on Total Scores

Comparing the total scores obtained in two countries for a specific test has generally been the only estimate of culture-fairness. If subjects in another country scored lower than those in the standardization sample, the test has been considered to be culturally weighted. Numerous studies are in this area; among them are those of Handel,[1] Thomas and Sjah,[2] and Vernon.[3]

It must be borne in mind, however, that total scores are influenced by sampling fluctuations. Stratified sampling across countries is always far from perfect because the criteria of stratification fluctuate from country to country.

Selection of Items by Use of Empirical Data from Different Countries

Conducting simultaneous tryouts of a preliminary test form in two countries yields empirical data for the selection of final items. Final items may be selected on the basis of similar item difficulty and discrimination data.

An example of this method is shown by Manuel[4,5] in the construction of the parallel English and Spanish editions of the Cooperative Inter-American Tests. The criteria used in item selection were indices of item difficulty and expression of the same thought with approximately the same number of words. The tests are for various levels from first grade through the first year in college. They consist of a test of scholastic ability yielding non-verbal, numerical, and verbal subscores, and a test of reading comprehension with vocabulary, comprehension, and reading speed subscores.

---

[1] A. Handel, The suitability of certain non-verbal tests for testing immigrants in Israel. J. educ. Res., 1957, 51, 55-58.

[2] R. M. Thomas, & A. Sjah, The Draw-a-Man Test in Indonesia. J. educ. Psychol., 1961, 52, 232-235.

[3] P. E. Vernon, Intellectual development in non-technological societies. In G. Neilson (Ed.), Proceedings of the XIV International Congress of Applied Psychology, Vol. 3. child and education. Copenhagen, Denmark: Munksgaard, 1962. Pp. 94-105.

[4] H. T. Manuel, The construction of interlanguage tests, Eighteenth yearbook, National Council on Measurement in Education, 1961. Pp. 101-105.

[5] H. T. Manuel, Testing the speed of reading by parallel tests in English and Spanish, Nineteenth yearbook, National Council on Measurement in Education, 1962, Pp. 5-9.

In a study exploring the reliability of the advanced Inter-American Reading Tests, Manuel[1] obtained .80 and .81 Kuder-Richardson reliabilities for the English version and .80 and .84 for the Spanish version in two samples of American college students studying second year Spanish. For these two samples, the correlation of scores on the English and Spanish versions were .61 and .57.

Selection of Content from Different National Sources and Comparison of Item Characteristics

Another approach to constructing international tests the content of which is not geared solely to one culture is represented by a recent study by the UNESCO Institute for Education.[2] Twelve countries where eight different languages were spoken participated in the study, testing thirteen-year-olds with a battery of five tests: Non-Verbal Aptitude, Mathematics, Reading Comprehension, Geography and Science. All the tests except the Non-Verbal Aptitude were developed jointly by representatives of the participating countries, most of the items being taken from English, French, German, Israeli and American sources. The preliminary forms of the Non-Verbal Aptitude Test were constructed by the National Foundation for Educational Research in England and Wales. The tests including verbal material were translated into the vernacular of each country.

---

[1] H. T. Manuel, The use of parallel tests in the study of foreign language teaching. Educ. psychol. Measmt., 1953, 13, 431-436.

[2] A. W. Foshay, et al., Educational achievements of thirteen-year-olds in twelve countries. Hamburg: UNESCO Institute for Education, 1962.

Several results reached by the UNESCO study are interesting especially because they illustrate the type of psychometric criteria which can be used in across-country evaluations of culture-fairness.

An analysis of variance compared within-country variations in scores with across-country variations. Compared to variations on other tests, differences between countries in reading comprehension were remarkably small. The variance between countries on the Reading Comprehension Test was only 6 per cent of the average within-country variance. The between-country variance of the presumably culture-fair Non-Verbal Aptitude Test was approximately twice as great. Of course, it must be admitted that the latter test originated from only one country.

An analysis to determine the extent to which individual test items retained their difficulty with administration in a different country and in a different language involved correlating the percentage of correct responses for each item between pairs of countries. The tests in reading and mathematics yielded the highest between-country correlations, an average of .87 in both cases. However, the correlations on the reading test had a smaller range, from .80 to .98 compared to a range of .60 to .98 on the test in mathematics. Highest correlations on the reading test were for countries where the same language was spoken, such as .98 between England and Scotland, and .96 between the United States and England.

When intercorrelations of item difficulties were submitted to a rotated factor analysis, it was observed that a general factor of difficulty accounted for most of the variance between countries on all tests. Loadings on the remaining factors contrasted countries of different language groups, such as factor 2 in the reading test which discriminated English-speaking from French-speaking countries.

Relationship of Methods Used to Present Study

The methods used by Manuel and by the UNESCO Institute for Education are preferable to a comparison of total scores since they provide more stringent psychometric criteria in the evaluation of culture-fairness.

Selecting content from different national sources to incorporate into the international test, as used by the UNESCO Institute for Education, assures that the test will contain some elements common to each culture. However, there is no guarantee of the culture-fairness of the complete test. In addition, face validity in selecting content as judged by the international committee may not work empirically.

Manuel's method of selecting content on the basis of simultaneous tryout in the respective countries assures the empirical validity of the final test. Yet the countries represented by the foreign students in the United States are so numerous that constructing pairs of equivalent tests for each country and the United States on the basis of tryouts might result in an unduly elaborate enterprise.

Furthermore, neither the method used by Manuel, nor the method used by the UNESCO Institute for Education show, from a psychometric standpoint, what types of shifts in test and item characteristics might be expected if the test were developed in only one country and, subsequent to translation, administered in another. Explorations should be made to determine if the net effect of these transformations is appreciably great. If it is not significant, the possibility of translating an American reading test into the vernacular of each group of foreign students might be attempted. These translated tests might perhaps be accepted as measures parallel to their English version for foreign students.

Shortcomings of Supplementary English Proficiency Measures as a Solution to the Language-Fairness Problem

To overcome partially the contamination of scores on American college level screening devices by the level of English proficiency, it has been suggested that English proficiency tests could be used as supplementary measures. Examples of such tests are the ECT (English Composition Test), the Michigan Test, and the TOEFL (Test of English as a Foreign Language). One could thus estimate the extent to which the student has mastered the English language and could make allowance for language handicap on the standardized verbal measure such as the SAT.[1]

---

[1] The College Entrance Examination Board and the Institute of International Education, U.S. college and university policies, practices, and problems in admitting foreign students. New York: Institute of International Education, 1965. P. 26.

However, supplementing verbal measures with English proficiency tests does not seem to meet the problem of scholastic aptitude measurement. What is not tapped by American verbal tests because of functioning in a foreign language cannot be replaced by language proficiency tests. This argument seems to pertain particularly to cases of standardized American measures which deal with a connected sequence of ideas, such as reading tests.

Low Predictive Validity for General Achievement

At the college level, English language proficiency tests are not good predictors of general academic achievement for foreign students. Tests used in assessing English language proficiency tend to have a large overlap in the factors they measure, as shown by their intercorrelations; yet they have low correlations with college grades.

Allen[1] found that for 59 undergraduates of non-English-speaking background the correlation between first semester grade point average and scores on the Michigan Test of English Language Proficiency was only .29. In another study of foreign students by Kaplan and Jones,[2] English proficiency tests had less predictive validity for college grade point average and ratings by faculty than a reading test geared for American students. The battery used in the study consisted of the

---

[1] W. P. Allen, International student achievement: English test scores related to first semester grades. Houston, Texas: Office of International Student Advisor, University of Houston, 1965. (Mimeographed.)

[2] R. B. Kaplan, & R. A. Jones, Evaluation of relative foreign student success. Language learning, 1965, 14, 161-166.

Brown-Carlson Listening Comprehension Test, the University of Southern California English Placement Test, the Larry-Ward Test of Articles and Particles, and the Advanced form of the California Reading Test. The students were also interviewed and were asked to write an English theme. Intercorrelations of scores on the different measures ranged from .34 to .80. The California Reading Test, the interview, and the Larry-Ward Test were given the highest *beta* weights in the multiple correlations with grade point average and composite rating by faculty. The two multiple correlations were only in the .30s. As it may be observed from the relative *beta* weights in the multiple correlations, even considering the problem of language handicap, scores on the California Reading Test assessed academic aptitude better than language proficiency tests.

Difference in Processes Involved in Various Levels of Foreign Language Learning and Fluency in the Vernacular

Success in the beginning and advanced levels of foreign language learning do not have the same requisites of mental ability. Bovee and Froehlich[1] correlated achievement in first and second year French as shown by performance on the Cooperative French Test with scores on the Stanford-Binet. The correlation between mental ability scores on the Stanford-Binet and achievement in first year French was negligible; a correlation of .18, which was similar to the result by Gardner and

---

[1] A. G. Bovee, & G. J. Froehlich, Some observations on the relationship between mental ability and achievement in French. *Sch. Rev.*, 1945, 53, 534-537.

Lambert[1] dealing with the relationship of intelligence to high school performance in French. However, the correlation between mental ability and achievement in second year French in the Bovee and Froehlich study was .59 implying that mental ability plays an important role in advanced level foreign language achievement.

It is interesting to observe that the correlation between relatively advanced level foreign language achievement and mental ability, as obtained by Bovee and Froehlich, approximates the correlation between mental ability and general scholastic achievement at the level of high school graduation and college entrance. Bloom and Peters[2] reviewed studies involving correlations between freshmen grades and scores on various aptitude tests taken at the terminal point of high school, and found that they had a median value of .45.

The similar relationships of scholastic achievement and advanced levels of foreign language achievement to general mental ability may be explained by the increasing role of verbal fluency in advanced foreign language competence. Lambert[3] found that different linguistic and literary skills differentiate degrees of bilingualism.

---

[1] R. C. Gardner, & W. E. Lambert, Language aptitude, intelligence, and second language achievement. J. educ. Psychol., 1965, 56, 191-199.

[2] B. S. Bloom, & F. R. Peters, The use of academic prediction scales for counseling and selecting college entrants. New York: The Free Press of Glencoe, 1961. Pp. 19-25.

[3] W. E. Lambert, Developmental aspects of second language acquisition. J. soc. Psychol., 1956, 43, 83-104.

Mills[1] correlated scores on fifteen prognostic tests and seven personality tests with the language proficiency ratings of students who had completed two years of French. The multiple correlation was in the .70s. A factor analysis showed that verbal fluency was the only important factor accounting for the common variance. Wittenborn and Larsen[2] confirmed the findings of Mills in a factor analytic study of advanced German achievement.

In foreign language reading tests which tap reasoning and verbal fluency a distinction is made between reading in the vernacular and in a second language. Lado[3] limited foreign language reading for students at an intermediate level to the understanding of contextual meanings and denotations of words known to all native speakers. In Lado's schema, paragraph organization, contextual inference of partial and extended meanings of words, and stylistic variations can be tested only with advanced students. Reading for literary appreciation is treated as a process exclusively peculiar to reading in the vernacular. Thus, in the treatment of reading given by Lado, there is a gradation of verbal processes which increase in their complexity along with competence in a foreign language and proceed to that level which exists in the vernacular.

---

[1] S. R. Mills, Prognostic tests of ability in modern languages. Unpublished doctoral dissertation, University of London, 1942.

[2] J. R. Wittenborn, & R. P. Larsen, A factorial study of achievement in college German. J. educ. Psychol., 1944, 35, 39-48.

[3] R. Lado, Language testing: the construction and use of foreign language tests. London: Longmans Green, 1961. Pp. 228-232.

Implications for Present Study

The above analysis suggests a contrast between the intellectual processes involved in mastering the structural characteristics of a foreign language, and processes such as inference and reading comprehension that are dealt with at more advanced stages of language learning. The low correlation of English proficiency test scores with scholastic achievement in college may be related to the fact that these tests do not deal with the latter. As already illustrated, even reading comprehension examinations designed for foreign students make restrictions on the intricacy and complexity of processes to be tested.

This restriction may mean that an important factor is being eliminated from reading tests for foreign students that may measure one of the most important elements of intellectual ability. Therefore, verbal measures uncontaminated by language handicap, perhaps reading tests in the vernacular of foreign students, may be the best predictors of college level academic achievement.

Problem Investigated by Present Study

As suggested by the foregoing discussion, the present study has a theoretical and a practical orientation.

Theoretical Query: Culture-Fairness

From a theoretical standpoint, the present study poses the following two questions:

1. To what extent does a college level American reading test transform its character and nature when translated into Turkish and

administered to Turkish students? As already shown, previous studies have either combined test adaptation with translation, or developed tests simultaneously in two countries. Therefore, the net effect of transformations in a test developed originally in another culture could not be studied.

2. How much do translation and language differences cause changes in the difficulty and character of the test in contrast to cultural differences in modes of response to the tasks employed? This analysis involves the following procedures: Administering the English and translated Turkish reading tests to students in the respective countries in the vernacular, and comparing the two sets of results which include cultural as well as language and translation effects. Retranslating the Turkish translations into English with subsequent administration to American students, whereby an analysis of original and re-translated versions could be made in order to detect only language and translation effects. There has been no previous attempt to compare the psychometric characteristics of original, translated, and re-translated versions of a test in order to detect the differential effects of variables dealing with translation and language as opposed to culture.

It must be acknowledged that results based on the American and Turkish cultures cannot be generalized to all intercultural comparisons. The determination of cultural and linguistic proximity falls within the realm of cultural anthropology. The comparisons of psychometric characteristics and the interrelationships between various statistical indices attempted in the present study explore the desirability of extending such comparisons to general evaluations of culture-fairness involving other languages and cultures.

Practical Query: Language-Fairness

The practical application of this enterprise is the possibility that a pair of equivalent forms of reading tests, one in English and the other in the native language of the examinee, may provide a powerful diagnostic tool in assessing foreign student aptitude. One could identify the two factors, academic aptitude and mastery of English, which, up to the present, have been concealed in a single verbal score.

1. The student's potential for higher education would be revealed by his reading ability in his own language.

2. The extent to which his academic potential is depressed by the necessity to function in a foreign language would be revealed by the approximation of his English reading score to that in the vernacular.

Providing native language measures would also meet the present demand for aptitude tests for foreign students who know very little English or no English at all.

The level of high school graduation and college entrance has been chosen for the present study, since most of international educational exchange takes place at this level.

Related Studies. A study in progress by Littrell, Opstead and

Hara[1] involves translating American aptitude and achievement tests into Japanese for use in screening Japanese students for American univerisites. The tests are being administered to university applicants in Japan and to Japanese students within the United States.

Chan[2] also explored the possibility of contrasting scores on an English and Chinese version of a reading test in the screening of Chinese students in the United States. Students' scores on the Chinese translation of a college level reading test were compared with scores on the original English version, yielding a correlation of .61 between ability to read in the two respective languages. Chinese reading passages were also translated into English and administered to American students. The American group found the passages of Chinese origin more difficult than those which were originally English. On the other hand, the Chinese students had been in the United States for some time and did not find the Chinese translations of English passages as difficult.

Assumptions

In proposing this analysis for evaluating foreign students' aptitude and mastery of English, it is assumed that the reading process is uniform regardless of the language in which it is conducted.

---

[1] R. T. Littrell, P. E. Opstead, & T. Hara, The effectiveness of native language tests in predicting relative academic success. In Research in international education: research in progress and research recently completed, 1964-1965 survey. New York: National Association for Foreign Student Affairs and the Institute of International Education, 1965, Pp. 20-21. (Abstract)

[2] Y. Chan, The development of parallel reading comprehension examinations in English and Chinese at the graduate level. Unpublished doctoral dissertation, Teachers College, Columbia University, 1953.

This is why the reading score in the vernacular can replace the score in the foreign language. Several studies lend support to this assumption:

1. Reading in the vernacular seems to be a relatively uniform process in terms of reading inference, as found by the UNESCO Institute for Education.[1]

2. Gray[2] found the basic mechanical processes involved in reading, such as eye movements, to be universally shared by many cultures.

3. Coffman[3] found objective-type reading comprehension questions to work better than other verbal items such as analogies and antonyms with students from countries where objective testing is not so widely spread.

However, this does not mean that one can actually obtain a perfect correlation between reading in two languages when one language is a foreign language for the students. One probably cannot get a much higher correlation than the .61 obtained by Chan.[4] Complete bilingualism is a very rare case. People vary in the length, type, and quality of second language instruction they receive; some take longer to master the structural characteristics of a language than others. Mastery of linguistic skills, as already shown, has little to do with grasp of literary and stylistic devices which are dealt with at advanced stages of language learning.

---

[1] A. W. Foshay, et al., Educational achievements of thirteen-year-olds in twelve countries. Hamburg: UNESCO Institute for Education, 1962.

[2] W. S. Gray, The teaching of reading and writing: an international survey. Paris: UNESCO, 1956.

[3] W. E. Coffman, Evidence of cultural factors in responses of African students to items in an American test of scholastic aptitude. Twentieth yearbook, National Council on Measurement in Education, 1963, 28-37.

[4] Y. Chan, The development of parallel reading comprehension examinations in English and Chinese at the graduate level. Unpublished doctoral dissertation, Teachers College, Columbia University, 1953.

Carroll[1] has remarked that there are individual differences in fluency in the vernacular and has considered foreign language achievement in terms of ability in the vernacular. Spencer[2] has also wondered whether English language proficiency could be measured without reference to native language proficiency. However, both writers were probably referring to a case where, once an individual reaches an optimum level of achievement or a limit of facility in a foreign language, he approximates his facility in his own language. These assumptions would not refer to the individuals in this study who do not have full competence in English.

Measures in a foreign language can be language-fair only if an individual has reached this limit of facility in a foreign language. In all other cases, parallel measures in the vernacular are more likely to be the best predictors of reading comprehension.

---

[1] J. B. Carroll, Problems of testing in language instruction: some principles of language testing. In A.A. Hill (Ed.), Report of the fourth annual roundtable meeting on linguistics and language teaching. Washington, D.C.: Georgetown University Press, 1953. Pp. 6-10.

[2] R. E. Spencer, An abstract of the results of the English Language Proficiency Tests for international students. University Park, Pa.: Office of Examination Services, The Pennsylvania State University, 1961. (Mimeographed)

CHAPTER II

DESIGN AND PROCEDURE

Several questions were explored concerning the equivalence of the original English, translated Turkish, and re-translated English versions of the reading tests.

1. To what extent do two parallel forms of the English tests retain their comparability in Turkish in terms of total score?

2. How similar are the English and Turkish versions of the tests in terms of relative difficulty of individual items?

3. What types of shifts in the sharpness of discrimination of specific items can be seen in the different versions of the tests?

4. Is there a drop in the reliabilities of the instruments when they are translated and administered in a different culture? Is there a difference in reliability between the original and re-translated versions of the tests administered in the same culture?

Instruments

The study used different versions of two parallel forms of a reading test, along with a practice test, a vocabulary test, a questionnaire, and supplementary validation measures consisting of SAT-Verbal scores, intelligence test scores and school grades.

International Reading Test

The instruments central to the study were two parallel forms of a reading comprehension test appropriate for high school graduates and college entrants, their versions translated into Turkish, and their versions re-translated from Turkish back into English.

Forms A and B, the parallel forms of the original English versions were developed on the basis of a pilot tryout in 1963-1964. The administration of the preliminary forms was conducted in three colleges in the United States. There were four preliminary forms of the tests, each consisting of four reading comprehension passages with eight to twelve objective-type items based on each passage. Each item contained five options. From indices of difficulty and discrimination, ten passages and six items for each passage were selected for use in the final versions of the tests. Each of the two final versions thus consisted of five reading passages and thirty items. The average index of difficulty for both forms was .69. The five passages in each form were arranged in order from easiest to most difficult; items for each passage were ordered in a similar way.

The content of the reading tests was expository, dealing mostly with the social sciences and the humanities. A brief description of the nature of the reading passages is given below:

Form A

Passage 1: Volunteer services in America
Passage 2: Precedent in law
Passage 3: The artist and his audience
Passage 4: Subject matter of economics
Passage 5: Growth of civilization

Form B

Passage 1: Need for understanding
Passage 2: Organized religions
Passage 3: Nechaev's philosophy
Passage 4: Child study
Passage 5: Role of observation in science

These English forms of the reading tests were translated into Turkish by the present writer with the assistance of several other native speakers of Turkish. Emphasis was placed upon making the translations as direct as possible, but also upon reflecting the various stylistic devices of English in the Turkish translations. This was possible in many cases with minor alterations in wording.

During May, 1964, as a preliminary try-out of the materials, 136 Turkish high school juniors and seniors were tested with the translated forms of the reading tests. The two alternate forms retained comparability in difficulty, with an average difficulty index of .44 and .42 for forms A and B. As may be observed from the lower difficulty indices than those obtained in the United States, the translated tests administered in Turkey appeared to be more difficult due either to the translations, to cultural differences, or to the lower grade level of the Turkish groups to which they had been administered.

A supplementary study was conducted both as a preliminary check on the accuracy of the translations and as a means of obtaining a statistical measure of shifts in item difficulty which translation itself might have produced. The translated Turkish tests were re-translated back into English by an independent translator, a Turkish expert who was a native speaker of English. Comparison of the original and the re-translated English texts resulted in minor revisions of the Turkish text where both translators agreed that significant changes in meaning had been produced. The re-translation of form A was called form C; of form B, form D.

Practice Test for Turkish Students Tested in the Vernacular

A practice test consisting of a reading passage with twelve reading comprehension items was prepared for use in coaching Turkish students prior to the actual test administration. This was one of the easier passages used in the pilot tryout and not included in the final versions.

Vocabulary Test for Turkish Students Studying English

The Vocabulary Test G-T,[1] Forms 1 and 2, each consisting of twenty 5-option items were used for Turkish students studying English who took one form of the reading test in English.

Questionnaire for Turkish Students Studying English

A short questionnaire was constructed for Turkish students studying English. It inquired about previous study of English, field of specialization, and the college or university in which the students were planning to study in the United States.

---

[1] Institute of Psychological Research, Teachers College, Columbia University, Vocabulary Test G-T. New York, Author, 1962.

Supplementary Measures

The following measures were used in the various phases of sampling, and the determination of the validity of the International Reading Test.

SAT-Verbal Scores. For American college students, Scholastic Aptitude Test Verbal scores were collected in order to characterize the sample in this study in comparison with the total group entering college. These scores were also used to ascertain the similarity of the students tested with the alternate forms in terms of scholastic aptitude.

Intelligence Test Scores. Intelligence test scores were collected for the American high school seniors in order to determine the comparability of the aptitude of the students tested with the alternate forms. Two of the three high schools used the Otis Intelligence Test; the third used the California Tests of Mental Maturity in about 80 per cent of the cases and various other measures such as the Otis, the Wechsler, the Stanford-Binet, and the Lorge-Thorndike for the remainder of the students.

University of Ankara Entrance Examination Scores. Scores on the University of Ankara entrance examination battery were used to give a rough estimate of the achievement level of the Turkish high schools included in the study. This battery of tests is one of the national screening devices at the terminal point of high school education in Turkey. In a report by Caliskaner and Ozgentas,[1] Turkish high schools had been ranked on the basis of average scores obtained by their students applying to the University of Ankara in September, 1963. This ranking on the basis of the average scores on the four subtests of aptitude and achievement was used as a guide in the selection of high schools in which testing was conducted.

Grades in Literature. For the one college and six of the nine high schools tested in Turkey, the average of literature grades for the previous academic year were collected. Grades in literature are based upon oral expression, composition, and readings from Turkish, Middle-Eastern, and Western writers all read in Turkish. For public high schools grades have a range from 1 to 10; the passing grade is 5. In the one private school tested letter grades A, B, C, D were being used.

First Semester Grades in an American University. Where at least one semester had lapsed since the testing of Turkish students in an English language program, these students were followed up by a questionnaire. Students were asked to indicate the courses taken and the grades received during their first semester in the university where they had started to study in their major fields of specialization.

---

[1] A. Caliskaner, & I. Ozgentas, Ankara Universitesi 1963 giris sinavlarinda liselerin basari dereceleri. Ankara: Milli Egitim Bakanligi, Talim ve Terbiye Dairesi, Test ve Arastirma Burosu, 1964.

Sample

Sampling within two countries with varying educational systems and educational selectivity is evidently a difficult matter. Under summary of school statistics between 1953 and 1957, the UNESCO World Survey of Education[1] reported that 25 per cent of the total population in the United States was enrolled in school, while for Turkey this percentage was 10 per cent. Increase in school enrollment and in the number of diplomas granted also differs between Turkey and the United States. During this interval, the number of high school diplomas granted in the United States had increased by 19 per cent; in Turkey this increase was 70 per cent.

An analysis of the total school-going population in the two countries and the increase in enrollment between 1953 and 1957 is shown in Table 1 by educational level. It may be observed that, of students in school, high school enrollment in Turkey was less than half the percentage of that in the United States, but the enrollment increase in the mid 1950s was much higher in Turkey. The same is also true at the college level.

Turkish High School Seniors

The Turkish high school sample consisted of 714 seniors from seven public high schools or lises, one experimental high school, and

---

[1] International Documents Service, UNESCO world survey of education III: secondary education. New York: 1961. Pp. 1094, 1103, 1363, 1378.

Table 1

Analysis of Total Population in School and Increase in Enrollment between 1953-1957 in the United States and Turkey

Educational Level	Population in School		Enrollment Increase	
	U.S. %	Turkey %	U.S. %	Turkey %
Pre-primary	5	x^a	x^a	x^a
Primary	67	86	17	29
Secondary	21	9	26	88
Vocational and technical	x^a	3	33	47
Teacher training colleges and universities	7	2	50^b	81

x^a = no data reported

b = only teachers colleges

one technical institute for girls. According to the UNESCO World Survey of Education,[1] there are 114 public lises in Turkey, two experimental schools, and eighty five technical institutes for girls. The Turkish lise is sponsored and directed by the Directorate-General of Secondary Education. It comprises the ninth, tenth, and eleventh grades. In the tenth grade students branch out into either the section for arts or the section for science based upon their aptitude and interest. Students receiving the lise diploma are entitled to enter a university faculty or any other institution of higher learning. The experimental high schools follow approximately the same curriculum as the lises, but have more flexibility in terms of course choice. Their examination and grading system is similar to that of the American Comprehensive School. The technical institutes, authorized by the Under-Secretariat of State for Vocational and Technical Education, go only through the tenth grade. Emphasis in the last two years is upon vocational and technical training. Students receiving a certificate may enter a teacher training school for elementary education, or they may continue their studies in a lise.

---

[1] International Documents Service, UNESCO world survey of education III: secondary education. New York: Author, 1961. Pp. 1098-1100.

For the present study, sampling within the Turkish lises was conducted by referring to a report which indicated the scores achieved by secondary school graduates who had taken the entrance examinations of the University of Ankara in September, 1963.[1] Students admitted are required to achieve a minimum score on this battery consisting of four subtests of intelligence, and general achievement in natural science, social science, and foreign language. Scanning this report, it was observed that schools in the vicinity of Ankara showed considerable variability in scores achieved. Although a sample controlling for factors such as geographic location and urban-rural background would have been preferred, selection from the metropolitan area considerably simplified the process of contacting the schools and arranging the testing schedule.

---

[1] A. Caliskaner, & I. Ozgentas, Ankara Universitesi 1963 giris sinavlarinda liselerin basari dereceleri. Ankara: Milli Egitim Bakanligi, Talim ve Terbiye Dairesi, Test ve Arastirma Burosu, 1964.

Turkish lises are either co-educational, all male, or all female. For purposes of equating the number of boys and girls in the sample to be tested, three schools of each type were selected.

One or two homerooms were tested in each school included in the sample. The Turkish public school system utilizes heterogeneous grouping within each science or arts section. To control for any influence the two different curricula might have upon test scores, where two homerooms were tested within a particular school, and where this seemed convenient, one homeroom was chosen from the arts section, the other from the science section.

Table 2 recapitulates the sampling data on the Turkish secondary school students tested for the present study. Ranks of the seven lises on the entrance examinations are shown in columns 1 through 5; school type is shown in column 6. Columns 7 through 10 show the number of students and the number and type of homerooms tested in each school.

Since the number of public and private schools represented in the examinations totaled to 163, higher achieving schools are somewhat over-represented in the present sample. However, it is a good cross-section of the top two thirds of the schools ranked on the basis of total scores on the criterion.

An approximately equal number of males and females were represented in the sample. Fifty-three per cent were males, 47 per cent females. The average age of the students tested was 18.46, with a standard deviation of 1.48.

Table 2

Turkish High School Sample with Respect to Rank on the Criterion, School Type, and Students Tested

School Name	Rank on Criterion[a] Total Score[b]	Subtests[c]				School	N	Homerooms Tested[d]		
		1	2	3	4	Type	Tested	1	2	3
Yenimahalle	11	13	8	14	36	Female	89	1	1	
Ataturk	26	23	27	42	41	Male	91	1	1	
Cumhuriyet	27	25	36	28	24	Co-ed	99	1	1	
Ankara	33	43	69	20	26	Female	97		2	
Bahcelievler Experimental	40	35	53	51	28	Co-ed	62	1	1	
Gazi	47	47	39	52	59	Male	78	1	1	
Kurtulus	86	76	80	89	93	Male	53		1	
Bayazit	92	102	60	93	95	Co-ed	103	1	1	
Bahcelievler Technical	x[e]	x[e]	x[e]	x[e]	x[e]	Female	42			1
Total							714	6	9	1

[a] Rank of average scores obtained by students within each school on the University of Ankara Entrance Examinations in September, 1963.

[b] Computed from (a) Intelligence, (b) Natural Science, (c) Social Science + Foreign Language scores.

[c] Column 1 = Intelligence; column 2 = Natural Science; column 3 = Social Science + Foreign Language; column 4 = Foreign Language. Items in the Foreign Language test were in English, French, or German. Students chose to take this test in the language in which they were most proficient. The ranks range from 1 to 163 and are based on standard scores for columns 1-3, raw scores for column 4.

[d] Column 1 = Science curriculum; column 2 = Arts curriculum; column 3 = Technical curriculum.

x[e] No scores on the criterion variable since this was a technical school.

Turkish College Students

Ninety-six first year students from the Middle East Technical University in Ankara comprised the Turkish college sample taking the tests in the vernacular. Instruction in this university established under the auspices of AID is in English. After being screened for and admitted to the university, students receive a year of English instruction before embarking upon their respective fields of study in the college proper. A group of these students were tested during the first month of this preparatory year.

About 70 per cent of the students tested were males; 30 per cent were females. The mean age was 18.84, with a standard deviation of 1.25.

American High School Seniors

A total of 587 high school seniors were tested in three high schools from the eastern United States. In the first high school, the original English versions, Forms A and B, were administered randomly among all students in the twelfth grade. Of these students, 180 were in the academic curriculum, 110 in the general curriculum. In the second and third high schools, only twelfth graders in the academic curriculum were tested. The re-translated English versions of the tests, forms C and D, were administered along with the original English versions A and B, the four forms being randomly assigned to students in a given class.

The equivalence of the aptitude level of the students who took the different forms and versions of the tests is shown in Tables 3 and 4.

Table 3

Means and Standard Deviations of Intelligence Scores for American High School Seniors Taking the Original English Versions

	Form A	Form B
Mean I.Q.	107.26	108.72
S. D.	10.34	10.44
N	179	189
No I.Q. information	15	15
N Tested	194	204

Table 4

Means and Standard Deviations of Intelligence Scores for American High School Seniors Taking the Original and Re-Translated English Versions

	Form A Original	Form C Re-translated A	Form B Original	Form D Re-translated B
Mean I.Q.	115.89	115.78	115.23	116.45
S. D.	10.93	10.89	11.89	9.56
N	46	51	47	51
No I.Q. information	7	5	3	3
N Tested	53	56	50	54

Most of the intelligence scores are based on the Otis Intelligence Test and the California Tests of Mental Maturity. Data in Table 4 are based upon the two schools in which all four forms of the tests were administered; those in Table 3 show results from administrations in all three high schools. It may be observed in these tables that the groups taking the tests in their different versions were equivalent with respect to aptitude.

Since approximately 25 per cent of all students taking forms A and B were in the general curriculum, an analysis of intelligence scores by type of high school curriculum is shown in Table 5. Considerable contrast in intelligence level is reflected in the two curriculum groups; but in each group students who took the two forms of the reading test did not differ significantly in intelligence.

Table 5

Means and Standard Deviations of Intelligence Scores for American High School Seniors in Academic and General Curricula

Curriculum		Form A	Form B
Academic	Mean I.Q.	112.56	113.52
	S.D.	10.91	10.95
	N	129	137
General	Mean I.Q.	93.80	96.06
	S.D.	10.77	13.19
	N	50	52

About 46 per cent of the American high school students tested were males; 54 per cent females. For the 61 per cent of students who had indicated their birth date, the mean age was 18.00, with a standard deviation of .49.

American College Students

The American college sample consisted of 816 students from four colleges in the eastern and southern United States. Three of the colleges offered a four-year liberal arts program; one was a two-year community college. In the former three colleges the two original English versions of the reading tests, forms A and B, were administered randomly to the students. The fourth college participated in the administration of both the original and the re-translated versions. Forms A, B, C, and D were administered randomly to students in a given class.

SAT-Verbal scores were available for most students in the four-year colleges. The two-year community college did not require the SAT for admission. Table 6 shows the comparability of students who had taken forms A and B with regard to aptitude on the SAT Verbal section.

Although the group taking form B seems to be a little higher in ability than the one taking form A, this difference is not statistically significant.

The present sample is fairly representative of the total group applying to colleges in the United States that require the CEEB. The average scores of students tested in spring, 1958 by the College

Table 6

Means and Standard Deviations of SAT-Verbal Scores for American College Students

	Form A	Form B
Mean SAT-Verbal	496.42	505.33
S. D.	93.87	96.98
N	280	270
No SAT Information	42	30
N Tested	322	300

Entrance Examination Board was 481 for males and 491 for females on the SAT-V.[1] According to norms based on scores received in the spring of 1956, boys in liberal arts enrolled in colleges which required both the SAT and Achievement tests had an average SAT score of 564, girls 571. Boys enrolled in colleges requiring the SAT only had a mean score of 513, girls 517.[2]

The American college students tested were mostly freshmen with a few upperclassmen who were in the psychology and education courses in which the tests were administered. About 57 per cent were males, 43 per cent were females. For the 88 per cent of students for whom age information was available, the mean age was 21.71, with a standard deviation of 6.41.

---

[1] College Entrance Examination Board, College Board score reports: A guide for counselors. Princeton, New Jersey: Author, 1958. P. 17.

[2] J. Fishman, 1957 supplement to College Board scores, No. 2. Princeton, New Jersey: College Entrance Examination Board, Educational Testing Service, 1957. Pp. 34-39.

The age distribution of American college students is shown in Table 7. The distribution is highly skewed. Most of the subjects are 17 to 24 years old. Evidently a few older students were enrolled in courses on a part-time basis.

Table 7

Age Distribution of American College Students

Age Interval	Frequency	Age Interval	Frequency
17-20	449	37-40	13
21-24	185	41-44	13
25-28	19	45-48	6
29-32	9	49-52	9
33-36	10	53-56	2
		Total	715
		No age information	101
		N Tested	816

Turkish Students Studying English

The sub-population of Turkish students studying English consisted of Turkish secondary school graduates and graduate students receiving special English instruction in preparation for studying in colleges where the curriculum is in English. At present, there are 1,070 Turkish students in the United States at various stages of study.[1]

Seventy-four Turkish students in the New York metropolitan area who were in the English language instruction programs of Queens College, New York University, Columbia University, and 12 students receiving the same kind of preparation at the Middle East Technical University in Ankara, Turkey were tested between September, 1964 and July, 1965. The sample constituted approximately 75 per cent of Turkish students in the respective schools in the United States who were at the intermediate or advanced levels of English instruction.

Most of the students in the United States were grantees of Turkish government scholarships or scholarships from private agencies in Turkey. Some had a minimal knowledge of English before coming to the United States; some did not know any English.

Of the 86 students who were tested, 41 took form A in English and form B in Turkish, and 45 took form B in English and form A in Turkish.

Seventy-nine students were males, 7 were females. Their age range was from 18 to 37, with a mean of 24.77 and a standard deviation of 4.48.

Administration Procedure

The following paragraphs outline the details of administration for the various groups in the study.

---

[1] Institute of International Education, Open Doors 1965: Report on International Exchange, New York: Author, 1965. Pp. 4-6.

Turkey

Each Turkish high school and college student in the sample took both forms of the reading test in Turkish. Since Turkish students are relatively unfamiliar with objective tests, and since practice and testing under power conditions have been found to decrease the influence of this condition to some degree,[1,2] students were given liberal time allowance. A 20-30 minute coaching with the practice test also preceded the actual administration. During this period, students were asked to read the passage silently; then the test administrator read the items based on the passage, answering each item, showing students how to read the stem and the options of each question, how to refer back to the reading passage to determine what was implied, and how to mark the correct option on the separate answer sheets.

A three-hour testing period was allowed for each administration. Test papers were collected when almost everybody in a given class had completed the tests. A short rest period was given after the administration of the first form.

Students were tested in groups ranging from twenty-five to a hundred. In each group testing order was counterbalanced among the students by having every other student take form A or B first.

---

[1] R. R. Knapp, The effects of time limits on the intelligence test performance of Mexican and American subjects. J. educ. Psychol., 1960, 51, 14-20.

[2] Majula Mukerjee, Effect of practice on test score. J. psychol. Res. Madras, 1962, 7, 37-42.

All the tests were administered in Turkey by the present writer, with the occasional assistance of one or two proctors where the group in a single administration exceeded sixty.

United States

Administration in the American high schools and colleges was carried out by sending the test materials to the respective schools, and asking professors or teachers to administer them in a 50-minute class period. Each student took only one form of the test, the different forms being randomly assigned among students in a given class. Test administrators were asked to instruct students to answer the questions according to what was stated or implied in each reading passage, not on the basis of general knowledge.

Turkish Students Studying English

Participation on the part of Turkish students receiving English instruction was voluntary, except for a group of twelve students from the Middle East Technical University in Turkey. With the cooperation of the directors of the American language programs, Turkish students were given a brief summary about the purpose and aims of the study, and were asked to participate in a 2 to 2.5 hour administration after their regular courses.

In each testing session, students first filled in the questionnaire; then took one form of the test in English, the other in Turkish. Testing order of the English-Turkish versions was counterbalanced, half the students taking each form first. In consecutive testing sessions, different forms of the tests were given in English and Turkish. Ten minutes was allowed for the vocabulary test which was administered after the reading tests.

The tests were administered by the present writer.

Students tested in 1964 and January, 1965 were followed up after a semester to determine the colleges and universities in which they were enrolled, the courses they had taken, and the grades they had received.

Analyses of Data

The data analyses consisted of scoring the tests, subsample analyses, reliability determination, and the application of correlational, variance, and item-analysis procedures.

Scoring

The tests were scored on the basis of items answered correctly. Since most students had attempted all the items, no problems were introduced with regard to a correction factor for unattempted items.

Analysis of Variance

Means and standard deviations were computed for the different forms and versions of the reading tests administered to the various groups. The same were also computed separately by order of administration for the Turkish forms administered in Turkey.

Four different analyses of variance tested for significance of the following differences in mean scores.

1. Test form and country differences for college students.

2. Test form and country differences for high school students.

3. Differences in the original and re-translated versions of the two English forms administered to American students.

4. Test form and order differences in the two alternate Turkish forms administered to Turkish students.

Subsample Analysis

The data for the Turkish high school and the American college groups were allocated into two subsamples equated on the basis of total score. Answer sheets for students in each school in the American sample were ordered with respect to total score, and alternate students were allocated into subsamples 1 and 2. Thus two stratified subsamples equated on total score were obtained for each college in the United States. A slight modification was made for the Turkish high school sample where each student took two forms of the test. Rankings of students on a particular form of the test was made for alternate schools, each student being allocated to one of the two subsamples. For example, answer sheets of School 1 were ranked on the basis of total score on form A; those for School 2 on the basis of total score on form B. A student in School 1 who had obtained the highest score on form A was placed into subsample 1 regardless of his score on form B. These subsamples were used to investigate the stability of item difficulty and discrimination indices.

Correlation of Item Difficulty Indices

An index of difficulty for each item was obtained by computing the percentage of students answering the item correctly. Since negligibly few students had not finished the tests, no correction factor was introduced for the last items.

Item difficulty indices for each form of the test were correlated within and between countries. A similar correlation analysis was applied to data from the original and re-translated English versions of the tests. Difficulty indices on the English versions by Turkish students studying English were also correlated with indices from American students taking the test in the vernacular.

The average across-country correlations were corrected for the unreliability of difficulty indices within a particular country as reflected by the correlation between subsamples within the country.

Since the samples in the re-translation analysis and the administration of the English tests to Turkish students were considerably smaller than the samples of American and Turkish groups on which the between-country correlations were based, a correction formula for sample size was applied in inferring the reliabilities for the former samples from the reliabilities of the latter. The Spearman-Brown formula $_n r_{xx} = \frac{n\, r_{xx}}{1 + (n-1)\, r_{xx}}$ was used, where r_{xx} is the reliability of the test as shown by the correlation of difficulty indices between subsamples of a country and n is the multiplicand for sample size. For example, the n for form B of the re-translated English version would be $\frac{83}{178} = .466$, since the sample taking the re-translated form B included 83 cases, and the reliability was originally computed on 178 cases. A correction for attenuation was then applied, using these adjusted reliabilities in the evaluation of a particular correlation.

Correlation of Item Discrimination Indices

Point biserial correlation coefficients of each item with total score were computed. These indices of discrimination were obtained for the right option, as well as the remaining four wrong options in each item of which most gave negative correlations. The coefficients for the right options were then corrected for the inclusion of that item response in the total score by the formula $r_i\ (t - i) = r_{it} - \frac{\sigma i}{\sigma t}$ where $\sigma_i = \sqrt{pq}$.[1]

The discrimination indices for correct options were correlated for subsamples within each country and across the two countries. The average across-country correlations were corrected for the unreliability of discrimination indices within one country as shown by the magnitude of the correlations within the two subsamples of the particular country.

Correlation of Difficulty and Discrimination Indices

The difficulty indices were correlated with discrimination indices within each country in order to see if there was a difference between the two countries in this respect.

---

[1] K. I. Howard, & G. A. Forehand, A method for correcting item-total correlations for the effect of relevant item inclusion. Educ. psychol. Measmt., 1962, 22, 731-735.

Correlation of Popularity of Errors

The popularity of the four wrong options in each item was of as much interest as the difficulty of the right option, since this might reflect cultural differences and any changes of meaning introduced by translation from English into Turkish and re-translation from Turkish into English.

The analysis of the relative popularity of specific errors was accomplished by adjusting the percentage remaining after choice of the right option to 100 per cent, and by considering the four remaining options as a percentage of this total. Correlations of the 120 percentages based on the thirty items in each form were computed for the two subsamples within one country and the subsamples across countries. The average of the across-country correlations were corrected for attenuation by use of the within-country correlations.

A similar correlational analysis was applied to percentages based on the re-translation study where the original and re-translated English versions were administered randomly among American students. Another correlation of this type was obtained from Turkish students taking the tests in English and American students taking the tests in their own language.

Evaluation of the Relative Across-Country Difficulty of Specific Items and Item Responses

The analysis of the relative across-country difficulty of specific items and item responses was applied to data from the American college and Turkish high school students. Taking the average item difficulty for the two countries as a reference, a range of twenty points was used in order to identify items which were especially easy or difficult for one country but not for the other. Twenty-nine of the 60 items in the two forms were thereby isolated. Fifteen were easier for Turkish students; 14 were easier for Americans.

Scanning the popularity of specific errors, one could also identify certain responses that were more popular among students of one country. When the percentage remaining after choice of the right option was adjusted to 100, a frequency distribution of the between-country differences in the percentages showed that two-thirds of these differences were below 15 per cent. A difference of 15 per cent was therefore chosen as the cut-off point to identify variations between the two countries in the popularity of specific errors.

For some items that were unusually difficult for Turkish students, and for some that were especially easy, no specific item response accounted for this shift in difficulty. In other cases, however, specific wrong options were more popular for one country, although no significant difference in overall item difficulty could be observed.

Items with differential difficulty also containing differences in the popularity of specific errors were isolated for further analysis, along with those items that contained popular wrong options for one country although the overall item difficulty remained the same. A review of the original English, translated Turkish, and re-translated English forms was made to see if the observed differences might be accounted for by any of several factors:

1. Changes in phrasing or stylistic variations.
2. Cultural differences.
3. Differential familiarity with taking objective tests.
4. Differential familiarity with the content of the reading passages, which might have resulted in choosing an option by general knowledge.

The discrimination indices of these items were also analyzed to note if relative item difficulty was associated with better discrimination. This was accomplished by considering the discrimination index of a right option in a particular country as a deviation from the mean discrimination index of all right options. From the pool of items that seemed easier or harder for one country, one could thus see if the specific item was among the better or poorer discriminating ones. The difference between the deviation scores for the United States and Turkey showed in which country the discrimination power of the item was more effective.

Relative Difficulty of Reading Passages

The average difficulty of the six items based on each reading passage was computed for American college and Turkish high school students. A rank correlation indicated whether or not the content of reading passages had a differential effect on item difficulties in the two countries.

Indices of Reliability

Since the reading tests used in the present study were new instruments, reliability data were obtained for interpretation of the research findings.

Internal Consistency. Kuder-Richardson Formula 20 reliabilities were computed for the original English, translated Turkish, and the retranslated English versions of the tests. This measure indicated any transformations in the general efficiency of the tests caused by translation and use in a different culture.

Stability of Difficulty and Discrimination Indices within One Country. Allocation of data for American college and Turkish high school students into two subsamples equated on total score permitted the evaluation of difficulty and discrimination indices in terms of the reliability of these measures within a particular country.

Alternate-Form Reliability. Correlating scores on the Turkish forms for Turkish high school and college students to whom both form A and B were administered constituted this index of reliability.

Alternate-Language Reliability. For Turkish students studying English, the correlation between scores on the English and Turkish versions was computed. This correlation was compared with the alternate-form reliability of the test in Turkish.

Correlation with Other Measures

Total scores for American college students were correlated with SAT-Verbal scores; scores for American high school students were correlated with scores on intelligence tests.

For Turkish high school and college students, scores on the reading tests were correlated with the average of last year's literature grades. A separate correlation coefficient was obtained for each of the seven schools for which grade data were collected.

From the follow-up questionnaire sent to Turkish students studying English, a qualitative evaluation of subsequent success in their respective fields became possible.

CHAPTER III

RESULTS AND CONCLUSIONS

The study yielded promising results with respect to the use of translated tests as relatively culture-fair measures of reading ability. Several observations also suggested possibilities for further experimental techniques to be employed in the development of international screening devices.

Total Test Scores

The two forms of the tests in English and Turkish were of approximately the same difficulty for American and Turkish samples in the same grade. Furthermore, the two alternate forms retained their comparable difficulty with translation and administration in a different culture. Table 8 summarizes the data based on total scores for high school and college samples in the two countries.

Since approximately 25 per cent of the American high school students tested were in the general curriculum, it may be appropriate to distinguish between the scores achieved by students in the academic and general curricula. Table 9 shows an analysis of this kind. The difference between students in the two curricula with respect to intelligence scores, already illustrated in Chapter II, was also reflected in the scores achieved on the reading tests. For both groups, however, the two forms retained their comparable level of difficulty. The variance of the scores was more restricted for students in the

Table 8

Means and Standard Deviations of Total Scores for High School and College Students in the United States and Turkey

Group		Form A		Form B	
		U.S.	Turkey	U.S.	Turkey
College	Mean Score	19.66	18.70	19.20	18.35
	S.D.	4.68	4.36	5.14	4.14
	N	381	96	356	96
H.S.	Mean Score	14.64	15.71	14.58	15.04
	S.D.	6.47	3.61	6.39	3.64
	N	194	714	204	714

Table 9

Analysis of the Means and Standard Deviations of Total Scores for American High School Students by Curriculum

Group		Form A	Form B
Academic	Mean Score	17.19	17.12
	S.D.	5.31	5.37
	N	139	149
General	Mean Score	8.68	8.79
	S.D.	3.78	3.84
	N	55	55

general curriculum.

Analyses of variance applied to total scores for each grade level and each test form in the United States and Turkey yielded non-significant differences as shown in Tables 10 and 11.

Table 10

Analysis of Variance of Total Scores on the Two Forms Obtained by American and Turkish College Students

Source of Variation	Degrees of Freedom	Mean Square	F	F .95
Between Means	3	56.67	2.47	2.61
Within Groups	925	22.98		

Table 11

Analysis of Variance of Total Scores on the Two Forms Obtained by American and Turkish High School Students

Source of Variation	Degrees of Freedom	Mean Square	F	F .95
Between Means	3	11.30	.58	2.60
Within Groups	1822	19.32		

Nor did there appear any striking shifts in the overall difficulty of the tests as they were translated into Turkish and re-translated back into English. Tables 12 and 13 show the average scores and standard deviations of the re-translation study, and the analysis of variance testing for translation and test form effects.

Table 12

Means and Standard Deviations of Total Scores for the Original and Re-Translated English Versions Administered to American Students

	Form A Original	Form C Re-translated	Form B Original	Form D Re-translated
Mean Score	16.98	15.65	17.02	16.71
S.D.	5.15	4.52	5.14	4.90
N	112	106	106	83

Table 13

Analysis of Variance of Total Scores on the Original and Re-Translated English Versions of the Two Parallel Forms

Source of Variation	Degrees of Freedom	Mean Square	F	F .95
Between Means	3	43.67	1.78	2.62
Within Groups	403	24.60		

Table 14 shows the means and standard deviations of scores achieved by Turkish students studying English on the English and Turkish versions of the reading test, and on the Vocabulary Test G-T. Each student took one form of the reading test in Turkish, the alternate form in English.

Table 14

Means and Standard Deviations of Total Scores for Turkish Students Studying English on the Reading Test and the Vocabulary Test G-T

	International Reading Test				Vocabulary Test G-T
	Form A		Form B		
	Turkish	English	Turkish	English	
Mean Score	18.24	11.40	17.42	11.91	7.00
S.D.	3.82	4.51	3.69	4.43	2.35
N	41	45	45	41	52

The results again show that the alternate forms of the reading tests retained their comparable difficulty. As might be expected, English scores were much lower than scores in Turkish. On the other hand, achievement on the Turkish versions did not differ from the achievement of the other Turkish and American college groups tested. English scores were lower than the scores of all other groups who took the test in the vernacular. The language handicap of these students, some of whom had not completed their period of English instruction, is definitely illustrated by the results.

As shown in Table 14, scores on the 20 word vocabulary test administered in English were also low. In the United States norming sample, subjects with 9 to 12 years of schooling had a mean score of 10.68, with a standard deviation of 2.55. Subjects with 13 or more years of education achieved a mean score of 13.74, with a standard deviation of 2.66.[1] Although the scores of Turkish students had about the same variance as the norming sample, the mean was considerably lower.

An analysis of variance for test form and testing order effects on the total scores of Turkish high school and college students is shown in Table 15. The order of taking the tests had a negligible influence on total scores achieved, but Form B was more difficult. The statistical significance of a total score difference of about .5 between the two forms may be partially due to the large number of cases used in the analysis and to the elimination of between-persons variance. The interaction between effects of test form and testing order, as it may be observed in Table 15, was literally zero.

Item Difficulty

The correlations of item difficulty indices between subsamples of American college and Turkish high school groups are indicated in Table 16, along with a correction of between-country correlations for the unreliability of the measures within a particular country.

Two observations may be made on the basis of Table 16. First, the correlations of difficulty indices within each country are almost unity. Second, across-country correlations of item difficulty are lower but still high. Correction for attenuation raises the across-country correlations to an average of .69 for the two forms, an

---

[1] Vocabulary Test G-T. Directions and Norms. Institute of Psychological Research, Teachers College, Columbia University, New York, 1962.

Table 15

Analysis of Variance of Total Scores on the Two Turkish Forms with Differential Testing Order

Order	Test Form Form A X	 Mean	 Form B X	 Mean	 Total X	 Mean
1	6405	16.17	5716	15.62	12121	15.91
2	5915	16.16	6156	15.54	12071	15.84
Total	12320	16.17	11872	15.58	24192	15.87

Source of Variation	Degrees of Freedom	Mean Square	F	F .95
Between Forms	1	132.00	8.87	3.84
Between Orders	1	2.00	.13	3.84
Interaction	1	0.00		
Within Groups	1520[a]	14.89	3.00	2.60

[a] No information on testing order = 96

increase of only one to two points.

The tests seem to be a little more difficult for Turkish high school seniors than for American college students. However, this 13 to 14 per cent difference in average item difficulty does not seem to influence the correlations to a great extent. When, in a supplementary analysis, average item difficulties for the two subsamples of the American college group were correlated with item difficulties for the

Table 16

Correlations of Item Difficulty Indices for American College and Turkish High School Subsamples

(N = 30 items)

	Form A				Form B			
	U.S.		Turkey		U.S.		Turkey	
	Samp. 1 (192)	Samp. 2 (189)	Samp. 3 (357)	Samp. 4 (357)	Samp. 1 (178)	Samp. 2 (178)	Samp. 3 (357)	Samp. 4 (357)
Mean Difficulty	65	65	52	52	64	64	50	50
Within Country	r_{12} = .98		r_{34} = .98		r_{12} = .97		r_{34} = .98	
Between Countries	r_{13} = .66		r_{14} = .68		r_{13} = .72		r_{14} = .67	
	r_{23} = .63		r_{24} = .66		r_{23} = .71		r_{24} = .66	
Between Countries Average			$\bar{r}$ = .66				$\bar{r}$ = .69	
Between Countries Average Corrected for Attenuation			r = .67				r = .71	

Note. 1. Subscripts of r refer to subsamples.

2. The numbers in parentheses are values of N in each subsample.

Turkish college sample, the correlation for Form A corrected for sample size and attenuation was .71, for Form B, .70. These indices are almost identical to the average correlations shown in Table 16. When average item difficulties for the two subsamples of the Turkish high school group were correlated with item difficulties for the Turkish college sample and then corrected for sample size and attenuation, the correlation for form A was .97 and for form B .95. This showed the stability of item difficulties across the two age and grade groups of a particular country.

Table 17 shows a similar correlation analysis between the original and re-translated English versions of the reading tests administered to American students. The correlation, corrected for sample size and attenuation, between the original and re-translated versions of form B seems to be somewhat lower than that for form A, .77 as against .95. Their average is .86 which is higher than the across-country correlations. Furthermore, there does not seem to be a noticeable difference in overall difficulty due to translations since average item difficulties are quite similar.

Table 18 shows the correlations between item difficulties of the English versions administered to Turkish students studying English and to American college students. Indices for the latter group have been averaged for the two subsamples. As may be observed, correction for sample size and attenuation increases these correlations from .79 and .71 to .83 and .77 respectively. This relationship is slightly higher than the degree of relationship obtained when the tests were administered in the vernacular to American and Turkish samples. It is slightly lower than the correspondence between item difficulties in the original and re-translated English versions. Considering the

Table 17

Correlations of Item Difficulty Indices for Original and Re-Translated English Versions Administered to American Students

(N = 30 items)

	Form A		Form B	
	Original (112)[a]	Re-translated (106)	Original (106)	Re-translated (83)
Mean Difficulty	56	52	57	56
Reliability[b] Corrected by Spearman-Brown Formula	.96	.96	.95	.93
Uncorrected	r = .91		r = .72	
Corrected for Attenuation	r = .95		r = .77	

[a] The numbers in parentheses are values of N in each sample.

[b] Reliability of difficulty indices as shown by the correlation between subsamples of a country.

language handicap of the Turkish students tested, one wonders whether this correlation reflects cultural differences more than variables caused by lack of proficiency in the language. Applying a similar analysis to subjects who are nearly bilinguals would shed more light on this issue.

Item Discrimination

Analyses relating to the discrimination indices of right options are shown in Table 19. The first row shows average point biserial correlation coefficients for each form of the test for the subsamples of

Table 18

Correlations of Item Difficulty Indices for English Versions Administered to American College Students and Turkish Students Studying English

(N = 30 items)

	Form A		Form B	
	U.S. (381)[a]	Turkey (45)	U.S. (356)	Turkey (41)
Mean Difficulty	65	38	64	40
Reliability[b] Corrected by Spearman-Brown Formula	.99	.91	.98	.87
Uncorrected	r = .79		r = .71	
Corrected for Attenuation	r = .83		r = .77	

[a] The numbers in parentheses are values of N in each sample.

[b] Reliability of difficulty indices as shown by the correlation between subsamples of a country.

American college and Turkish high school students. Intercorrelations within and between countries are indicated in the lower rows.

It may be observed that the sharpness of discrimination of single items drops appreciably as the tests are translated into Turkish and administered in Turkey. However, the within-country correlations of the discrimination indices show considerable stability, averaging .66 for Turkey, and .52 for the United States on the two test forms. On the other hand, the between-country correlations are very low. The average is .07 for form A, and .10 for form B. Corrected for the unreliability of the discrimination indices within each country, these

Table 19

Correlations of Item Discrimination Indices for American College and Turkish High School Subsamples

(N = 30 items)

	Form A				Form B			
	U.S.		Turkey		U.S.		Turkey	
	Samp. 1 (192)	Samp. 2 (189)	Samp. 3 (357)	Samp. 4 (357)	Samp. 1 (178)	Samp. 2 (178)	Samp. 3 (357)	Samp. 4 (357)
Mean r_{pb}	28	26	15	13	30	29	14	13
Within Country	$r_{12} = .56$		$r_{34} = .80$		$r_{12} = .47$		$r_{34} = .51$	
Between Countries	$r_{13} = .17$		$r_{14} = .18$		$r_{13} = -.09$		$r_{14} = .22$	
	$r_{23} = -.02$		$r_{24} = -.06$		$r_{23} = .22$		$r_{24} = .06$	
Between Countries Average			$\bar{r} = .07$				$\bar{r} = .10$	
Between Countries Average Corrected for Attenuation			$r = .10$				$r = .20$	

Note. 1. Subscripts of r refer to subsamples.

2. The numbers in parentheses are values of N in each subsample.

correlations rise to .10 and .20 respectively.

The lack of similarity between countries in item discrimination power seems interesting, especially in view of the stability of item difficulty and discrimination characteristics within each of the two countries and the relative cross-cultural homogeneity in item difficulty.

Table 20 shows a supplementary analysis of the relationship between indices of difficulty and discrimination within the United States as compared to Turkey. The difficulty index for each item was averaged for the two subsamples and correlated with its average discrimination index in the subsamples. It is interesting to observe the contrast between the United States and Turkey in this respect. While there seems to be no correspondence between difficulty and discrimination within the United States, as shown by the average correlation of .03 for the two forms of the tests, in Turkey easier items have higher discrimination. The average correlation for the two Turkish forms of the tests is .47.

Table 20

Within-Country Correlations between Difficulty and Discrimination Indices

(N = 30 items)

Country	Form A	Form B
United States	.19 (381)	- .12 (356)
Turkey	.64 (714)	.30 (714)

Note. The numbers in parentheses are values of N in each sample.

Popularity of Errors

Tables 21 and 22 show the correlations of the popularity of wrong options for each item between the various groups tested with the different test versions. In Table 21 an analysis has been made pertaining to the within- and between-country stability of the index for each error obtained by adjusting the percentage remaining after choice of the right option to 100. Table 22 shows a similar correlation analysis between the original and re-translated English versions, and between American and Turkish samples who took the original English versions.

Although the within-country stability of this index dealing with errors, as shown in Table 21, is lower than that dealing with correct options, as shown in Table 16, it is still quite high. The average within-country correlation of .92 in Table 21 compares with the average of .98 in Table 16. However, the correlation between errors drops to an average of .37 across subsamples of the two countries, a degree of relationship lower than the .69 for correct options already shown in Table 16.

If one were to contrast these sets of correlations in terms of common variance, a larger difference would be observed. The correlation of .92 accounts for 85 per cent of the variance, the correlation of .98 accounts for 96 per cent of the variance. The across-country average correlation for the wrong options reflects 14 per cent of the variance as against 48 per cent in the case of right options.

The correlation of errors in the original and re-translated

Table 21

Correlations of Errors for American College and Turkish High School Subsamples

(N = 120 options)

	Form A		Form B	
Within Country	U. S.	$r_{12} = .87$	$r_{12} = .92$	
	Turkey	$r_{34} = .94$	$r_{34} = .95$	
Between Countries	$r_{13} = .46$	$r_{14} = .45$	$r_{13} = .36$	$r_{14} = .26$
	$r_{23} = .42$	$r_{24} = .41$	$r_{23} = .30$	$r_{24} = .28$
Between Countries Average		$\bar{r} = .44$	$\bar{r} = .30$	
Between Countries Average Corrected for Attenuation		$r = .49$	$r = .32$	

Note. Subscripts of r refer to subsamples

Table 22

Correlations of Errors for Original and Re-Translated English Versions Administered to American and Turkish Samples

(N = 120 options)

Group	Form A	Form B
American Students Taking Original and Re-Translated Versions	.70	.74
American and Turkish Students Taking Original Versions	.30	.44

English versions shown in Table 22 falls between the within- and between-country correlations, averaging .72 for forms A and B. It is also interesting to note by referring to Table 17, that the correlations of wrong options between the original and re-translated versions do not differ greatly from similar correlations of right options. These observations seem to imply that translation or language effects, although one cannot distinguish between them in the present study, transform item characteristics to a certain extent, but that cultural differences influence this transformation to as great or even a greater extent. This hypothesis is also partly supported by Table 22 showing the correlation between American and Turkish groups in the popularity of errors for tests taken in English. The average correlations of .30 and .44 between Turkish and American students taking the tests in English are almost identical to the correlations obtained when the tests were administered in the vernacular to the two groups.

Another implication of the findings, based on the cross-cultural stability of right options in contrast to the wrong ones, is that misleading item choices seem to differ more in their popularity for different cultural groups. What seems like a popular mislead to an item writer may not be so popular in a foreign culture.

Specific Items and Item Responses

Comparisons between relative difficulty and discrimination for all items that differed markedly in relative difficulty for the cultural groups are presented in Table 23. As outlined in Chapter II, a

Table 23

Relationship between Relative Item Difficulty and Discrimination for American College and Turkish High School Samples

Easier Items for U.S.						Easier Items for Turkey					
Form	Difficulty		Discrimination			Form	Difficulty		Discrimination		
Item	U.S. 1	Turkey 2	U.S. 3	Turkey 4	Differ. 5	Item	U.S. 1	Turkey 2	U.S. 3	Turkey 4	Differ. 5
A 13	76	34	- 02	- 09	+ 07	A 1	92	90	- 13	+ 12	+ 25
A 16	76	52	+ 15	+ 04	+ 11	A 2	74	88	- 10	+ 02	+ 12
A 17	58	30	+ 05	+ 01	+ 04	A 4	85	87	+ 05	+ 06	+ 01
A 18	67	41	+ 02	- 01	+ 03	A 8	78	81	+ 09	+ 04	- 05
A 19	70	30	+ 03	- 12	+ 15	A 14	72	82	+ 15	+ 05	- 10
A 21	78	42	- 09	- 04	- 05	A 24	24	38	- 01	- 02	- 01
A 26	70	34	+ 01	- 08	+ 09	A 29	26	44	- 24	- 11	+ 13
B 2	74	39	- 08	- 09	+ 01	A 30	20	32	- 10	+ 04	+ 14
B 3	77	38	+ 02	- 02	+ 04	B 5	57	58	- 07	+ 02	+ 09
B 4	50	26	+ 14	- 03	+ 17	B 6	42	40	+ 12	+ 01	- 11
B 9	60	24	- 02	- 12	+ 10	B 8	71	70	+ 04	- 04	- 08
B 14	93	62	- 09	- 02	- 07	B 10	58	66	- 05	00	+ 05
B 21	75	33	+ 10	- 10	+ 20	B 16	78	88	- 06	- 04	+ 02
B 25	80	56	- 02	+ 02	- 04	B 18	26	33	- 14	00	+ 14
						B 23	54	56	- 14	- 01	+ 13
Mean					+ 6.07						+ 4.86

twenty-point item difficulty range (after correcting for difference in means) was used for isolating these twenty-nine items. Columns 1 and 2 in each sub-table list item difficulties for each country. Columns 3 and 4 deal with discrimination indices expressed as deviations from the mean of the particular country. In each sub-table, column 5 indicates the difference between the deviation discrimination indices of the two countries. A + sign means that the item which was easier for the particular country also had better discrimination.

It may be observed in Table 23 that there is a positive relationship between relative item easiness and relative item discrimination power. This is true for both the United States and Turkey. Of the 14 items which were easier for American students, only 3 had poorer discrimination in the United States than in Turkey; of the 15 items which were easier for Turkish students, only 5 were poorer discriminators.

According to Table 23, the degree to which an item differentiates between high and low scorers in a particular country seems to be associated with its relative overall facility. However, this effect may arise solely because of the correlation in Turkey between item difficulty and discrimination, already shown in Table 20.

A review of all test items in terms of the relative cross-cultural and inter-language popularity of right and wrong options suggested that the items could be grouped into four different types. Two examples of each type are shown in Tables 24 - 27 which include indices of difficulty and discrimination for the various versions of the tests.

Table 24

Examples of Items with No Overall Difficulty or Error Difference

Form Item		Option A	B	C	D	E	Turkey: Adjusted Difficulty
A 9	Difficulty						
	Original English	02 (06)	10 (32)	15 (46)	04 (13)	66	
	Translated Turkish	04 (10)	10 (28)	20 (50)	04 (08)	62	75
	Re-translated English	02 (04)	19 (39)	20 (41)	08 (16)	51	
	Discrimination						
	Original English	-26	-24	-18	-06	44	
	Translated Turkish	-16	-04	-21	-10	30	
B 13	Difficulty						
	Original English	06 (27)	06 (23)	77	10 (44)	01 (04)	
	Translated Turkish	10 (32)	05 (16)	70	11 (38)	02 (08)	83
	Re-translated English	03 (13)	06 (26)	77	11 (48)	03 (13)	
	Discrimination						
	Original English	-16	-21	45	-26	-15	
	Translated Turkish	-14	-06	24	-11	-09	

Notes referring to Tables 24-27.

1. Numbers in parentheses are values obtained by adjusting the percentage remaining after choice of the right option to 100.

2. Underlined numbers indicate right answers.

3. The adjustment in difficulty for the translated Turkish versions shown in the last column was calculated by adding 13 points.

Table 25

Examples of Items with No Overall Difficulty Difference Containing Different Popular Errors

Form				Option			Turkey: Adjusted Difficulty
Item		A	B	C	D	E	
A 5	Difficulty						
	Original English	78	10 (46)	05 (22)	01 (04)	06 (26)	
	Translated Turkish	60	04 (09)	30 (76)	02 (07)	02 (07)	73
	Re-translated English	66	10 (29)	14 (41)	01 (03)	08 (24)	
	Discrimination						
	Original English	32	-22	-06	-03	-20	
	Translated Turkish	35	-10	-26	-11	-10	
B 19	Difficulty						
	Original English	02 (10)	84	11 (70)	01 (07)	02 (12)	
	Translated Turkish	18 (52)	65	08 (22)	02 (08)	06 (15)	78
	Re-translated English	03 (12)	76	18 (75)	01 (04)	00 (00)	
	Discrimination						
	Original English	-10	34	-24	-16	-09	
	Translated Turkish	-20	30	-14	-12	-04	

Table 26

Examples of Items with Overall Difficulty Difference Containing No Different Popular Errors

Form				Option			Turkey: Adjusted Difficulty
Item		A	B	C	D	E	
A 13	Difficulty						
	Original English	09	76	02	05	06	
		(38)		(08)	(20)	(28)	
	Translated Turkish	25	34	05	10	24	47
		(38)		(08)	(14)	(36)	
	Re-translated English	10	62	04	07	15	
		(26)		(10)	(18)	(39)	
	Discrimination						
	Original English	-24	34	-14	-06	-12	
	Translated Turkish	-08	18	-10	-05	-03	
B 5	Difficulty						
	Original English	22	57	12	04	04	
		(52)		(29)	(08)	(10)	
	Translated Turkish	23	58	06	03	10	71
		(54)		(16)	(07)	(23)	
	Re-translated English	15	58	05	08	13	
		(36)		(12)	(19)	(31)	
	Discrimination						
	Original English	-01	28	-12	-08	-20	
	Translated Turkish	-14	28	-14	-10	-10	

Table 27

Examples of Items with Overall Difficulty Difference also Containing Different Popular Errors

Form Item		A	B	C	D	E	Turkey: Adjusted Difficulty
A 30	Difficulty						
	Original English	20	10 (13)	13 (16)	44 (56)	08 (11)	
	Translated Turkish	32	20 (29)	08 (11)	22 (32)	10 (16)	45
	Re-translated English	15	11 (13)	04 (05)	03 (04)	55 (65)	
	Discrimination						
	Original English	26	-09	02	-10	00	
	Translated Turkish	30	-08	-06	-08	-10	
B 4	Difficulty						
	Original English	16 (32)	08 (14)	10 (20)	16 (31)	50	
	Translated Turkish	26 (34)	33 (45)	08 (11)	06 (08)	26	39
	Re-translated English	12 (27)	08 (18)	18 (40)	05 (11)	55	
	Discrimination						
	Original English	-34	-12	-16	-16	54	
	Translated Turkish	-06	-07	-10	-06	22	

1. Items with no overall difficulty or specific error differences. Five of the 60 items in the test were of this type. In Table 24 it may be observed that both difficulty and discrimination indices remain relatively stable throughout the different versions. Items of this kind seem to be the most desirable for cross-cultural evaluations.

2. Items with no overall difficulty difference containing different popular specific errors. Twenty-six of the total 60 items were of this kind. Examples are shown in Table 25. Although the difficulty of right options remained quite stable with translation and administration in a different culture, specific misleading options changed, to some degree, the nature of the item. In item A 5, options B and E were more popular in the United States than in Turkey; option C was more popular in Turkey. The popularity of these options in the re-translated English versions fell somewhere in between the American and Turkish profiles, implying that perhaps differences in the phrasing of the texts accounted for this variation.

On the other hand, it may be observed in item B 19 that although option A was more popular in Turkey and option C was more popular in the United States, the profile of the re-translated version is almost identical to the American profile. In this case this shift was evidently cultural.

For both items of this kind, greater popularity of a mislead seems to be associated with better discrimination power as shown by the higher negative r. This is true for both countries under comparison.

3. Items with overall difficulty difference containing no

different popular specific errors. Table 26 shows examples of two items which, although they were relatively more difficult for one of the two countries, had no specific misleading option that could account for this difference. An explanation of this phenomenon is hard to find, except for the conjecture that familiarity with the concepts in the reading passages might cause the differential effect. Of the 60 items in the test, 16 were of this type.

4. Items with overall difficulty difference also containing different popular specific errors. Thirteen of the 60 items in the test were of this type. Such items are the least culture-fair measures, since the difficulty of right answers is affected, at least in part, by the perceived suitability of specific wrong options in different cultures. For example in both items in Table 27 options B and D yielded different results in a Turkish culture and in an American culture. Looking at data from the re-translated versions, one can not reach a general conclusion as to whether these differences are due to culture or translation. Only in item B 4 are the percentages for the options more similar to the original in the re-translated English version than in the Turkish version. The discrimination indices for the options do not follow a general pattern either.

When the original and re-translated English and the Turkish versions of the tests were reviewed to see if the differences encountered were due to culture or translation, it was found that changes in phrasing between the two English versions were surprisingly few. Reviewers familiar with the Turkish culture attributed the variations mostly to cultural differences and to possible differential familiarity with the content of the reading passages.

Reading Passages

A general comparison of the extent to which groups of items based on a specific reading passage retained their relative difficulty when administered to Turkish students in the vernacular is shown in Table 28. The rank correlation of .92 of reading passage difficulties implies that the difficulties of items were not determined by the nature of the reading passage on which they were based. This correlation may also imply that the content of the reading passages retained their difficulty in the Turkish culture. It must be admitted, however, that the rank correlation is based on only a small number of cases.

Reliability

Table 29 shows the internal consistency reliabilities and the standard errors of measurement for the various versions of the reading tests. The reliability indices have been obtained by the Kuder-Richardson Formula 20. There seems to be a significant drop in the reliability of the tests as they are administered in a different culture and in a different language. The average correlation of .78 for the original English versions decreased to about .54 for the Turkish versions administered in Turkey. The re-translated versions yielded an average reliability of .74, a value quite similar to that of the original tests.

However, it might be worthwhile to note the difference in the

Table 28

Correlation of Difficulty of Reading Passages for American College and Turkish High School Samples

Reading Passage	U.S. Difficulty	U.S. Rank	Turkey Difficulty	Turkey Rank
Form A				
1	80	1	72	1
2	73	2	62	2
3	72	3.5	50	5.0
4	54	9	38	9.5
5	48	10	38	9.5
Form B				
1	64	6	46	7
2	58	8	48	6
3	66	5	58	3
4	72	3.5	57	4
5	60	7	41	8

$\rho = .92$

Table 29

Internal Consistency of the Original and Re-Translated English and Translated Turkish Versions

	Form A Original				Form C Re-translated A	Form B Original				Form D Re-translated B
	U.S.		Turkey			U.S.		Turkey		
	Samp. 1	Samp. 2	Samp. 3	Samp. 4	U.S.	Samp. 1	Samp. 2	Samp. 3	Samp. 4	U.S.
K-R$_{20}$	.78	.76	.57	.52	.72	.80	.80	.54	.50	.77
SEM	2.23	2.29	2.46	2.44	2.40	2.34	2.30	2.50	2.53	2.35
Mean	19.67	19.64	15.52	15.89	15.65	19.17	19.24	14.99	15.11	16.71
S.D.	4.75	4.67	3.73	3.53	4.52	5.19	5.10	3.68	3.56	4.90
N	192	189	357	357	106	178	178	357	357	83

variability of scores in the two cultural groups. The standard deviations of scores in the Turkish groups were smaller. It may well be that differences in the homogeneity of the groups tested caused the differences in reliability. In fact, the standard error of measurement of the Turkish tests are similar to the English forms although the reliability coefficients are not as high.

A follow-up of the variance in scores of the different grade groups tested showed that the Turkish college sample taking form A had a variance similar to the American groups. The Kuder-Richardson Formula 20 reliability of the Turkish form A for this group was .73, with a standard error of measurement of 2.27. This value of internal consistency approximates the reliabilities obtained for the American samples, and suggests that the drop in reliability in the Turkish tests can be accounted for by the homogeneity of the Turkish samples.

The alternate-form reliability of the instrument for Turkish students is presented in Table 30. For students tested only in Turkish, the alternate-form reliability is somewhat lower than the internal consistency reliability. The correlation between the English and Turkish forms for Turkish students who have been studying English long enough to be testable with the English versions is as high as the other two measures of reliability. This implies that one can get as good a prediction of these Turkish students' ability to read from a test of reading English as one can from another form of the test in Turkish.

For Turkish students studying English, scores on the Vocabulary Test G-T correlated with English reading ability to the extent of .51.

Table 30

Alternate-Form and Alternate-Language Reliabilities for Turkish Samples

Group		Form A		Form B	
		Turkish a	English b	Turkish c	English d
Students Tested in Turkish	Mean	16.06		15.46	
	S.D.	3.85		3.84	
	N	810		810	
			$r_{ac} = .47$		
Students Tested in English and Turkish	Mean	18.24	11.40	17.42	11.91
	S.D.	3.82	4.51	3.69	4.43
	N	41	45	45	41
			$r_{ad} = .49$		$r_{bc} = .58$

Correlation with Other Measures

The correlation of scores on the original and re-translated English versions of the reading tests with other aptitude tests is shown in Table 31. The external criteria were SAT-Verbal scores for college, and intelligence quotients for high school students. Since the junior college in which the re-translated forms were administered along with the original forms did not require the SAT for admission, correlations of the re-translations with external criteria could not be obtained at the college level.

Table 31

Correlation of Scores on the Original and Re-translated English Versions with SAT-Verbal Scores and IQ

Criterion	Form A Original	Form C Re-translated A	Form B Original	Form D Re-translated B
College SAT-Verbal	.55 (280)		.69 (270)	
High School I.Q.	.76 (179)	.59 (51)	.72 (189)	.60 (51)

Note. The numbers in parentheses are values of N.

Both the original and the re-translated English versions have high concurrent validity, as shown in Table 31. The process of translation and re-translation seems to have reduced these validity coefficients to some degree, showing a drop from an average .74 to an average of .60 for the two alternate forms. However, these lower correlations may have been due to excluding non-academic students and reducing the range for the group tested with the re-translated versions.

Table 32 presents Turkish reading score correlations with the average of literature grades for the preceding academic year. Data are for Turkish high school and college students. Separate correlations have been obtained for each school and each test form. It may be observed that the correlations show considerable range, from zero to .43. Variability is also observed for each school in the difference between the correlation for form A and form B. In one case this difference is

Table 32

Correlation of Scores on the Translated Turkish Versions with Grades in Literature

School Name	N	Correlation Form A	Correlation Form B
Yenimahalle H.S.	77	.43	.13
Ataturk H.S.	90	.02	.14
Bahcelievler Experimental H.S.[a]	62	.07	.40
Kurtulus H.S.	53	.53	.29
Bayazit H.S.	103	.07	.00
Bahcelievler Technical	41	.10	.02
Middle East Technical Univ.	93	.16	.33
	$\bar{r}$	.16	.19

[a] Letter grades were given in this school. All other schools used number grades ranging from 1 to 10, 5 being the passing grade.

more than 30 points. In general, these indices are much lower than the internal consistency or the alternate form reliabilities of the same tests. They are also lower than the concurrent validity indices of the English forms administered to American students. Of course, the external criteria used for the United States and Turkey are not directly comparable since correlations with school grades are generally lower than correlations with standardized tests. In a previous study in Jordan, composite scores of a battery correlated with teacher grades to the extent of .27, a value not very different from the correlations obtained in the present study.[1]

Turkish students studying English who were tested with the English-Turkish versions during 1964 and January, 1965 were followed up in the universities they had subsequently attended. Out of the approximately thirty students, ten had already enrolled in programs of their field of specialization, eight were still studying in American language instruction programs, and four were not studying at all for reasons such as illness or being engaged in independent research. The remaining eight had either moved from their past residence leaving no further address, or did not respond to the questionnaire.

In the questionnaire, students were asked to indicate the courses they had taken during the preceding semester, and the grades they had received. The data can only be subjected to a qualitative evaluation, since the number of cases was small and their grades were too uniform for a meaningful analysis.

Six of the ten students registered in a university proper were graduates, four were undergraduates. The universities in which they were studying were in the mid-west, the south, and the east. In general, they were doing well in their respective fields although they were not fully competent in English. Students placed in freshman English or composition courses along with American college students did not do so well in these courses, probably because of the competition they received from students whose native language was English. Students still studying English were also doing acceptable work in their programs.

---

[1] A. Adas, Patterns of achievement in Jordanian schools. Unpublished doctoral dissertation, Teachers College, Columbia University, New York, 1963. P. 112.

CHAPTER IV
DISCUSSION

The suitability of translated reading comprehension tests for cross-cultural evaluations was supported by the results of several analyses in the present research.

First, the total scores of American and Turkish students at similar educational levels were approximately the same when they took the tests in their own language. Differences between the two countries in curriculum and in the selection of students should be acknowledged, nevertheless, in interpreting the findings. Turkish students have completed eleven years of schooling by the time they graduate from the lise; American high school graduates have had twelve years of instruction. The subdivision into academic and general curricula does not exist in the Turkish lise, resulting in greater scholastic homogenity in Turkey at the terminal point of high school. Yet, overall reading achievement in the vernacular at the level of high school graduation and college entrance, as found in the present study, shows a close parallelism in the United States and Turkey.

Second, comparisons dealing with the relative difficulty of individual reading comprehension items showed their stability with translation and administration in the Turkish culture. Although the correlation between item difficulties was somewhat lower than similar correlations obtained in a previous study by the UNESCO Institute for

Education,[1] it may well be remembered that the tests used in the UNESCO study were jointly developed by representatives from various countries, selecting items and passages from different national sources. A query of the present study was whether reading passages and test items originally developed and selected on the basis of a pilot tryout in one culture would retain their characteristics with translation and administration in another culture. In this respect, the present research differs also from the already mentioned cross-cultural studies which dealt with the administration of standardized foreign tests after quite extensive transformation of their nature by local adaptation of some items and elimination of certain other items. From a theoretical point of view, the present study implies that careful translations of college-level tests measure what Holmes[2] called the ability to infer from a text, a process which is relatively uniform throughout cultures.

The study also showed that language differences between English and Turkish and the process of translation account for variations in text and test characteristics. Although the .85 correlation between item difficulties in the original and re-translated versions is very high, it still is not unity. Therefore, cultural differences between the United States and Turkey are not the only determiners of the shifts which are represented in the .70 correlation between item difficulties in the English and Turkish versions. One can speculate that if translations and language differences had not produced some alterations in meaning, the across-country correlations of item difficulties would have been even higher.

---

[1] A. W. Foshay, et al., Educational achievements of thirteen-year-olds in twelve countries. Hamburg: UNESCO Institute for Education, 1962.

[2] J. A. Holmes, Factors underlying major reading disabilities at the college level. Genet. Psychol. Monogr., 1954, 49, 3-95.

What part of the shifts observed in the re-translation study is due to language differences between English and Turkish, i.e., the lack of conceptual correlates, and what part to the skill of translators cannot be determined from the present data. It is true that in the Whorfian[1] thesis linguistic differences produce conceptual and behavioral differences which, in turn, may cause cultural disparity. However, the communication process in translations may also account for some of these shifts, as shown in studies by Posthuma,[2] and Miller and Beebe-Center.[3] Isolating the differential effects of these two factors would perhaps involve having the items with difficulty difference rated by translators, linguists, and cultural anthropologists.

It must also be remembered that the re-translation study compared tests that underwent the translation process twice: first from English into Turkish, and then, from Turkish back into English. This process might have tended to magnify effects due to the translation process. Furthermore, where translators' accuracy caused differences in text, it becomes difficult to know whether the variations occurred in going from English to Turkish or from Turkish back into English. Considering these possible sources of variance, the high correlation of .85 between item difficulties in the original and re-translated English versions suggests that the nature of reading content remains relatively uniform after translation into a different language and re-translation back into the original language.

---

[1] B. Whorf, Science and linguistics. In H. B. Allen (Ed.), Readings in applied English linguistics. New York: Appleton-Century Crofts, 1958. P. 33.

[2] F. J. Posthuma, Taalkundige aantekeningenbij een test voor verbale inprenting. Ned. Tijdschr. Psychol., 1954, 9, 526-534.

[3] G. A. Miller, & J. G. Beebe-Center, Some psychological methods for evaluating the quality of translations. Mechanical translation. 1956, 3, 73-80.

It is interesting to note that item difficulty, as reflected in response to the right option, is much more stable across cultures than are responses to the wrong or misleading options. This could perhaps mean that in cases where the text is not read carefully students resort to answering the item on the basis of general knowledge and, in turn, conclude with the wrong answer. A supplementary study was conducted after the termination of the testing as a follow-up of this idea. American graduate students were asked to answer the questions in the original English reading tests without reading the passages on which they were based. Item difficulties as well as responses to the wrong options were correlated for the tests administered with the reading passage and without reference to the passage. Responses to the right options correlated low in the .20s; responses to wrong options correlated about .50. Considering the fact that wrong option choices are somewhat influenced by responses to the right option, the latter correlation seems quite high and implies that general knowledge partly determines wrong option choices.

Furthermore, critical reading and paragraph comprehension are developmental skills, as pointed out by many writers such as Strang,[1] Smith,[2] and Rogers.[3] Emphases placed on types of reading skills such as reading for details and grasping the main idea are not uniform throughout the world. Revues of international comparisons of this kind have been given by Schmidt,[4] and the National Foundation for Educational Research in England and Wales.[5] Curricular emphases also produce differential familiarity in specific content areas. All these factors may contribute to the differences in the popularity of specific errors encountered in the present study.

Cultural differences in response to wrong test options do not seem to affect the nature of the right option. For example, in the present research, a right option generally retained its difficulty in the United States and Turkey, although in the same item one or two wrong options were chosen more frequently in one country than in the other. A future study may be suggested to eliminate the influence of cross-cultural differences in the popularity of wrong options upon the right answer in the few cases where this seems to be true. Items may be written for a pilot tryout in the two countries containing more options than the number which will eventually be used. The choices that seem to reflect cultural differences may be subsequently eliminated in the final form of the test.

---

[1] Ruth Strang, Diagnostic teaching of reading. New York: McGraw Hill, 1964. P. 215.

[2] D. E. P. Smith, Reading comprehension: a proposed model. Ninth yearbook, National Reading Conference for Colleges and Adults. Fort Worth, Texas 1960. Pp. 21-27.

[3] Bernice Rogers, Directed and undirected critical reading responses of high school students. Unpublished doctoral dissertation, University of Chicago, 1960. P. 209.

[4] B. Schmidt, A glance at developmental reading outside the United States. In A. J. Figurel (Ed.), Reading as an intellectual activity. New York: Scholastic Magazines, 1963. Pp. 215-216.

[5] National Foundation for Educational Research in England and Wales, in association with the Institute of Education, London University. Spelling irregularity and reading difficulty in English. London: Information Service of the National Foundation for Educational Research in England and Wales, 1960. P. 8.

One of the most crucial findings of the study is the contrast between difficulty and discrimination data used in cross-cultural comparisons. While item difficulties remained quite stable in the two cultural groups, item discrimination indices showed large shifts. The average correlation of .15 between the sharpness of discrimination of individual items in the United States and Turkey is almost negligible. Still, this does not mean that within the Turkish culture test items do not discriminate between high and low scorers. Although in general the sharpness of discrimination is lower, the stability of discrimination indices within Turkey is about .70, a value higher than the average of .50 for the United States.

Another finding of interest is the difference between the United States and Turkey in the correlation of difficulty and discrimination indices within each country. The finding of a relationship between item easiness and better discrimination in Turkey but not in the United States may be related to the fact that the test was originally developed in English and for an American culture. When a test of this kind is administered in a different culture, items which are culturally loaded probably appear to be more difficult. In turn, other psychometric functions such as indices of discrimination may be affected, thereby producing a positive relationship between difficulty and discrimination.

The above argument may also apply to the finding concerning relative item easiness and relative item discrimination across cultures. Culture-fair items may well be those that do not appear especially difficult for the country under comparison. As illustrated in this analysis, relatively easy and relatively difficult items were selected after equating average item difficulties in the United States and Turkey. Items relatively easier for the United States might have been those which, for the Turkish group, were culturally loaded and thus lost their discrimination power. Items which were easier for Turkey might have been the culture-fair items which also had better discrimination in Turkey.

Still another implication of the present study comes from the relationship between group heterogeneity and indices such as internal consistency, alternate-form reliability, and discrimination. The supplementary analysis done with the Turkish college group showed that controlling for restriction in group heterogeneity raised the internal consistency reliabilities of the Turkish tests, and made them similar to the reliabilities of the English tests. The relative homogeneity of the Turkish group may also explain the lower discrimination indices and the lower alternate-form reliability of the Turkish tests. In international testing one continuously deals with differences in the selectivity of students at different educational levels. Therefore, variations in group homogeneity should be taken into consideration before interpreting psychometric data dealing with reliability and discrimination.

For Turkish students studying English the .54 correlation between reading in Turkish and English, even with tests the reliability of which did not exceed this correlation, seems too significant to be neglected. Another factor which probably tended to lower this correlation is the fact that the Turkish and English versions were parallel forms and not identical tests. Even more important, however, is the variability of the group tested in English proficiency.

Several suggestions may be made for the perfection of the instruments to be used in the screening process for foreign students, and for the more general theoretical analysis of cross-cultural item statistics.

1. A test concurrently prepared in two languages, selecting reading passages and items in both countries might produce an effective culture-fair test. A preliminary test of this kind would have more items for each passage, and perhaps, more options for each item than the number generally used for a pilot tryout. Selections of the final items and options would be based on comparisons of data concurrently obtained in the two countries.

2. A test originally prepared in Turkish on the basis of item statistics from a pilot tryout in Turkey might be translated into English for administration in the United States. The resulting data might

be compared with those of the present study to see if the positive relationship between difficulty and discrimination is a general phenomenon which exists in all tests originally constructed in a different language and culture.

3. Cultural differences in test and item response in contrast to language differences might be studied by administering the English forms of the reading test in other English-speaking countries such as England, Australia, India, or Nigeria.

4. The extent to which the concepts expressed in the reading passages, rather than their specific wording, determine the difficulty of reading comprehension tasks may be analyzed by editing the original reading passages and by using the same items with texts of varying complexity in wording. Experimenting at the elementary school level, Wilson[1] found that concepts that were more difficult in the original versions tended to retain their difficulty after simplifications in wording. A similar study at the college level in a specific country or across countries might yield interesting results in terms of the suitability of translations where the same ideas are expressed differently.

5. Cultural differences in responding to items merely on the basis of general knowledge might be studied by testing native speakers in the two cultures only by the items in the reading tests. The subjects would answer the questions without reading the passage on which

---

[1] Mary C. Wilson, The effect of amplifying material upon comprehension. J. exp. Educ., 1944, 13, 5-8.

they were based. The analysis would have a twofold value, the first being the elimination of items which could be answered correctly on the basis of general knowledge; and the second, the detection of specific errors produced by cultural variations. In fact, a study by Preston[1] showed that American students did better than chance in a similar passageless reading test. The results were attributed to test-sophistication on the part of students where irrelevant, too broad, and too narrow choices were eliminated in order to arrive at the correct option. Differential performance by two cultural groups on a test of this kind would also show the effect of acquaintance with objective tests upon performance.

6. Modifications may be made in test format to find any influence that a multiple choice question might have upon a reading comprehension score in countries where students are not familiar with objective tests. Dressel and Schmidt[2] experimented in this area with American students, asking them to mark more than one option in certain cases, and in other cases to indicate the degree of certainty of their choice. According to the authors, some multiple-choice items involve one right answer, and others, a best answer. The Kuder-Richardson reliabilities of the tests in different format ranged from .67 to .78 and did not differ significantly from each other. On the other hand, an analysis of variance applied to test scores obtained under different experimental

---

[1] R. C. Preston, Ability of students to identify correct responses before reading. J. educ. Res., 1964, 58, 181-183.

[2] P. Dressel, & J. Schmidt, Some modifications of the multiple-choice item. Educ. psychol. Measmt., 1953, 13, 574-595.

conditions yielded significant results. The study may imply that cross-cultural differences might be partly attributed to test format.

7. A relatively culture-fair test developed concurrently in two countries might be translated into various other languages and administered in the respective countries to determine if it would be more resistant to shifts in reliability and discrimination power than a test which had its origin in one language and culture.

8. When a reliable instrument in the vernacular of specific groups of foreign students is developed with a parallel form in English, comparisons could be made to determine how the validity of the reading test in the vernacular compares with the validity of the English test in predicting grades in the United States. A multiple correlation analysis could be applied using the two tests in combination to see if still better prediction could be obtained.

It would be worthwhile to conclude by saying that future studies in the field of cross-cultural testing should be directed not toward a simple comparison of total scores, but to comparisons of responses to specific passages, to specific items, and even to specific wrong options. The interrelationship between different item statistics, such as those of difficulty and discrimination, also need further research since they may be indices of varying elements in varying cultures.

CHAPTER V

SUMMARY

The purpose of the present research was to study the comparability of an English college-level reading comprehension test to its Turkish translation administered in Turkey, and to its English re-translation administered in the United States. This comparison would yield data on the relative culture-fairness of translations of reading tests for a better evaluation of the verbal ability of foreign students planning to study in the United States. Testing with two equivalent forms of reading tests, one in English, and one in the native language of the examinee might provide a powerful diagnostic tool for identifying (a) the individual's potential for higher education, and (b) the extent to which this potential is depressed as the student is faced with the necessity of shifting from his native language to English as a language of instruction.

The instruments central to the study were two parallel forms of a reading test suitable for the level of college entrance in the United States. These tests were translated into Turkish, and re-translated from Turkish back into English. Supplementary measures employed were SAT-Verbal scores, intelligence scores, a vocabulary test score, a questionnaire, and school grades.

The sample consisted of 714 Turkish high school seniors, and 96 Turkish college students tested in Turkey with the Turkish versions; 398 American high school seniors, and 737 American college students tested with the original English versions; 189 American college students and high school seniors tested with the re-translated English versions; and 86 Turkish students studying English tested with one form of the reading test in Turkish, the other in English.

The following conclusions were reached on the basis of the present study.

1. Average total scores for Turkish and American students were quite similar for samples in the same grade.

2. Re-translation of the English test did not seem to alter its characteristics for American students in terms of total score.

3. The equivalence of the two parallel forms remained stable with translation and administration in the Turkish culture, and with re-translation into English.

4. A correlation of about .70 was obtained in the relative difficulty of single items in Turkish and in English. The same type of correlation between the original and re-translated English versions was about .85, suggesting that both translation and cultural differences influence indices of item difficulty.

5. Correlating discrimination indices for single items between the English and Turkish versions showed a negligibly small relationship of about .15 between the sharpness of discrimination of a specific item in the two cultures. However, discrimination indices were stable within each country.

6. Due to the greater homogeneity in the Turkish group, an appreciable drop was found in the internal consistency reliability of

the Turkish test and in the sharpness of discrimination of single items.

7. Items which seemed relatively easier for either country also had better discrimination power in the particular country. Item difficulty and discrimination correlated negligibly within the United States, but significantly in Turkey.

8. Responses to the wrong options of each multiple-choice item were stable within each country, but correlated only about .40 across cultures. Translations did not seem to have as much effect in changing responses to wrong options, since similar correlations between the original and re-translated English versions was .72.

9. Rank correlation of the difficulty of reading passages in the United States and Turkey was .92, a value higher than the correlation of specific item difficulties.

10. For Turkish students studying English, reading ability in the two languages correlated to the extent of .54, implying the existence of common reading factors.

References

Adas, A. Patterns of achievement in Jordanian schools. Unpublished doctoral dissertation, Teachers College, Columbia University, New York, 1963.

Allen, W. P. International student achievement: English test scores related to first semester grades. Houston, Texas: Office of International Student Advisor, University of Houston, 1965. (Mimeographed)

Anastasi, Anne. Psychological testing. New York: Macmillan, 1959.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. Differential Aptitude Test manual. New York: Psychological Corporation, 1959.

Bloom, B. S., & Peters, F. R. The use of academic prediction scales for counseling and selecting college entrants. New York: The Free Press of Glencoe, 1961.

Bolton, F. B. Value of several intelligence tests for predicting scholastic achievement. J. educ. Res., 1947, 41, 133-138.

Bovee, A. G., & Froehlich, G. J. Some observations on the relationship between mental ability and achievement in French. Sch. Rev., 1945, 53, 534-537.

Caliskaner, A., & Ozgentas, I. Ankara Universitest 1963 giris sinavlarinda liselerin basari dereceleri. Ankara: Milli Egitim Bakanligi, Talim ve Terbiye Dairesi, Test ve Arastirma Burosu, 1964.

Carroll, J. B. Problems of testing in language instruction: some principles of language testing. In A. A. Hill (Ed.), Report of the fourth annual round table meeting on linguistics and language teaching. Washington, D. C.: Georgetown University Press, 1953. Pp. 6-10.

Carroll, J. B. The study of language. Cambridge, Mass.: Harvard University Press, 1953.

Center for Applied Linguistics of the Modern Language Association of America. Testing the English proficiency of foreign students. Washington, D. C.: Author, 1961.

Chan, Y. The development of parallel reading comprehension examinations in English and Chinese at the graduate level. Unpublished doctoral dissertation, Teachers College, Columbia University, 1953.

Choudhuri, P. K., & Majumdar, P. K. Factorial approach to the problem whether verbal intelligence tests can be replaced by performance type intelligence tests. Indian J. Psychol., 1963, 38, 125-128.

Church, A. M. The standardization program summary report 1947. Hawaii educ. Rev., 1947, 36, 53-56.

Coffman, W. E. Evidence of cultural factors in responses of African students to items in an American test of scholastic aptitude. Twentieth yearbook, National Council on Measurement in Education. 1963, 28-37.

College Entrance Examination Board, College board score reports: a guide for counselors. Princeton, New Jersey: Author, 1958.

The College Entrance Examination Board and the Institute of International Education. U.S. college and university policies, practices, and problems in admitting foreign students. New York: Institute of International Education, 1965.

Committee on Educational Interchange Policy. College and university programs of academic exchange--suggestions for the study of exchanges of students, faculty and short-term visitors. New York: Author, 1960.

Committee on the Foreign Student in American Colleges and Universities. The college, the university and the foreign student. New York: Author, 1963.

Davis, Allison. Social class influences upon learning. Cambridge, Mass.: Harvard University Press, 1955.

Dressel, P., & Schmidt, J. Some modifications of the multiple-choice item. Educ. psychol. Measmt., 1953, 13, 574-595.

Eaton, M. T. A survey of the language arts achievement of sixth grade children in 18 counties and 6 cities in India. Res. Bull. Ind. Dep. publ. Instruct., 1942, 3, 1-75.

Fishman, J. 1957 supplement to College Board scores, No. 2. Princeton, New Jersey: College Entrance Examination Board, Educational Testing Service, 1957.

Foshay, A. W., et al. Educational achievements of thirteen-year-olds in twelve countries. Hamburg: UNESCO Institute for Education, 1962.

Gardner, R. C., & Lambert, W. E. Language aptitude, intelligence, and second language achievement. J. educ. Psychol., 1965, 56, 191-199.

Garth, T. R., Elson, T. H., & M. M. Morton. The administration of non-language intelligence tests to Mexicans. J. abn. soc. Psychol., 1936, 31, 53-58.

Gates, I. A. The correlation of achievement in school subjects with intelligence tests and other variables. J. educ. Psychol., 1922, 13, 129-139, 223-235, 277-285.

Gray, W. S. The teaching of reading and writing: an international survey. Paris: UNESCO, 1956.

Handel, A. The suitability of certain non-verbal tests for testing immigrants in Israel. J. educ. Res., 1957, 51, 55-58.

Harris, D. P. A survey of English language requirements and facilities for foreign students in United States institutions of higher learning. 1961. New York: National Association for Foreign Student Affairs, 1962.

Herskovits, M. J. Cultural anthropology. New York: Alfred A. Knopf, 1960.

Holmes, J. A. Factors underlying major disabilities at the college level. Genet. Psychol. Monogr., 1954, 49, 3-95.

Howard, K. I., & Forehand, G. A. A method for correcting item-total correlations for the effect of relevant item inclusion. Educ. psychol. Measmt., 1962, 22, 731-735.

Institute of International Education. Open doors, 1965: report on international exchange. New York: Author, 1965.

Institute of Psychological Research, Teachers College, Columbia University, Vocabulary Test G-T. New York: Author, 1962.

International Documents Service. UNESCO world survey of education III: secondary education. New York: Author, 1961.

Kamat, V. V. A revision of the Binet scale for Indian children. Brit. J. educ. Psychol., 1934, 4, 296-309.

Kaplan, R. B., and Jones, R. A. Evaluation of relative foreign student success. *Language learning*, 1965, 14, 161-166.

Keehn, J. D., & Protho, E. T. Non-verbal tests as predictors of academic success in Lebanon. *Educ. psychol. Measmt.*, 1955, 15, 495-498

Knapp, R. R. The effects of time limits on the intelligence test performance of Mexican and American subjects. *J. educ. Psychol.*, 1960, 51, 14-20.

Kretch, D., Crutchfield, R. S., & Ballachey, E. L. *Individual in society*. New York: McGraw Hill, 1962.

Lado, R. *Language testing: the construction and use of foreign language tests*. London: Longmans Green, 1961.

Lambert, W. E. Developmental aspects of second language acquisition. *J. soc. Psychol.*, 1956, 43, 83-104.

Littrell, R. T., Opstead, P. E., & Hara, T. The effectiveness of native language tests in predicting relative academic success. In *Research in international education: research in progress and research recently completed, 1964-1965 survey*. New York: National Association for Foreign Student Affairs and the Institute of International Education, 1965. Pp. 20-21. (Abstract)

Lorge, I., & Thorndike, R. L. *The Lorge-Thorndike Intelligence tests. technical manual*. Boston: Houghton-Mifflin, 1962.

MacArthur, R. S., & Elley, W. B. The reduction of socioeconomic bias in intelligence testing. *Brit. J. educ. Psychol.*, 1963, 33, 107-119.

McKillop, Anne, & Yoloye, E. A. The reading of university students. *Teach. Educ.*, 1962, 3, 93-107.

Malin, A. J. An Indian adaptation of the WISC. *J. voc. educ. Psychol.*, 1964, 10, 128-131.

Manuel, H. T. The use of parallel tests in the study of foreign language teaching. *Educ. psychol. Measmt.*, 1953, 13, 431-436.

Manuel, H. T. The construction of interlanguage tests. *Eighteenth yearbook, National Council on Measurement in Education*, 1961. Pp. 101-105.

Manuel, H. T. Testing the speed of reading by parallel tests in English and Spanish. *Nineteenth yearbook, National Council on Measurement in Education*, 1962. Pp. 5-9.

Miller, G. A., & Beebe-Center, J. G. Some psychological methods for evaluating the quality of translations. Mechanical translation, 1956, 3, 73-80.

Mills, S. R. Prognostic tests of ability in modern languages. Unpublished doctoral dissertation, University of London, 1942.

Mukerjee, Majula. Effect of practice on test score. J. psychol. Res. Madras, 1962, 7, 37-42.

National Foundation for Educational Research in England and Wales. Spelling irregularity and reading difficulty in English. London: Author, 1960.

Ombredane, A. Etude du comportement intellectuel des noirs congolais. Psychol. franc., 1957, 1, 19.

Pasricha, P., & Pagedar, R. M. Adaptation of "WAIS" to the Gujarati population. J. voc. educ. Guidance, 1963, 9, 174-184.

Pieter, J. Intelligence quotient and environment. Kwart. Psychol., 1939, 11, 265-322.

Posthuma, F. J. Taalkundige aantekeningen bij een test voor verbale inprenting. Ned. Tijdschr. Psychol., 1954, 9, 526-534.

Preston, R. C. Ability of students to identify correct responses before reading. J. educ. Res., 1964, 58, 181-183.

Rogers, Bernice. Directed and undirected critical reading responses of high school students. Unpublished doctoral dissertation, University of Chicago, 1960.

Schmidt, R. A glance at developmental reading outside the United States. In A. J. Figurel (Ed.), Reading as an intellectual activity. New York: Scholastic Magazines, 1963. Pp. 215-216.

Sexton, Patricia C. Education and income. New York: Viking Press, 1961.

Smith, D. E. P. Reading comprehension: a proposed model. Ninth yearbook, National Reading Conference for Colleges and Adults. Pp. 21-27.

Spencer, R. E. An abstract of the results of the English Language Proficiency Tests for international students. University Park, Pa.: Office of Examination Services, the Pennsylvania State University, 1961. (Mimeographed)

Stevanovic, B. P. The development of the child's intelligence and the Beograd revision of the Binet-Simon scale: summary data and results. Bull. Acad. Lettr. serbe, 1935, 1, 89-114.

Strang, Ruth. Diagnostic teaching of reading. New York: McGraw Hill, 1964.

Thomas, R. M., & Sjah, A. The Draw-a-Man Test in Indonesia. J. educ. Psychol., 1961, 52, 232-235.

Verhaegen, P. Utilite actuelle des tests pour l'etude psychologique des autochtones congolais. Rev. Psychol. appl., 1956, 6, 139-151.

Vernon, P. E. Intellectual development in non-technological societies. In C. Neilson (Ed.), Proceedings of the XIV International Congress of Applied Psychology, Vol. 3. Child and education. Copenhagen, Denmark: Munksgaard, 1962. Pp. 94-105.

Westbrook, C. H. The use of English group intelligence tests with Chinese students. Education, Univ. Shanghai, 1940, 3, 83-95.

Whorf, B. Science and linguistics. In H. B. Allen (Ed.), Readings in applied English linguistics. New York: Appleton-Century Crofts, 1958. Pp. 28-38.

Wilson, Mary C. The effect of amplifying material upon comprehension. J. exp. Educ., 1944, 13, 5-8.

Wittenborn, J. R., & Larsen, R. P. A factorial study of achievement in college German. J. educ. Psychol., 1944, 35, 39-48.

Wu, T. M. On the second revision of the Chinese Binet-Simon scale. Shanghai: Commercial Press, 1936.

Appendix

Sample Passage

(Original)

Precedent has a very important theoretical weight in most legal systems, and in all legal systems it has an important practical weight. There are those legal systems which purport to be based on certain abstract principles of justice. The Roman law and its descendants, which indeed constitute the greater part of the law of the European continent, belong to this class. There are other systems like that of the English law, in which it is openly stated that precedent is the main business of legal thought. In either case, no new legal term has a completely secure meaning until it and its limitations have been determined in practice; and this is a matter of precedent. To fly in the face of a decision which has been made in an already existing case is to attack the uniqueness of the interpretation of the legal language and is, in the fact itself, a cause of indeterminateness and very probably of consequent injustice. Every case decided should advance the definition of the legal terms involved in a manner consistent with past decisions, and it should lead naturally on to new ones. Every piece of phraseology should be tested by the custom of the place and of the field of human activity to which it is relevant. The judges, those to whom is confided the task of the interpretation of the law, should perform their function in such a spirit that if Judge A is replaced by Judge B, the exchange cannot be expected to make a mterial change in the court's interpretation of customs and of statutes. This naturally must remain to some extent an ideal rather than a thing already done; but unless we are close to the followers of these ideals, we shall have chaos, and what is worse, a no-man's land in which dishonest men prey on the differences in possible interpretation of the statutes.

When precedent is given proper weight, if Judge A were replaced by Judge B one should expect that

A. Judge B would have to consult with Judge A before rendering an opinion.
B. Judge B would have to interpret the law in his own way in order to ensure justice.
C. any agreement in the legal interpretations of the two judges would be only coincidental.
D. There would be a material difference in the way in which Judge B would interpret the law.
E. Judge B would interpret the law very much as would Judge A.

Abstract principles of justice are an important basis of

A. only the English law.
B. only the Roman law.
C. both English and Roman law.
D. only the descendants of the Roman law.
E. both the Roman law and its descendants.

In practice, precedent has an important weight

A. primarily in the Roman legal system.
B. only in the English law.
C. in legal systems based on abstract principles.
D. in all legal systems except the Roman.
E. in all systems of law, without exception.

The meaning of a new legal term can only become clear after the term has been

A. related to abstract principles of justice.
B. carefully defined by a legislative body.
C. involved in a number of legal decisions in different cases.
D. re-interpreted by two or more judges.
E. uniquely interpreted in legal language.

Precedent has its basis in

A. previous legal decisions.
B. legal theory.
C. judges' preferences.
D. indeterminateness.
E. Roman law.

In addition to the statute on which it was based, the meaning of legal phraseology should reflect the

A. abstract principles of justice.
B. wisdom of the judge in charge of the case.
C. established usage of the times.
D. uniqueness of the interpretation of legal language.
E. indeterminacy of human affairs.

Sample Passage

(Re-translated)

Precedent plays a very important theoretical role in many legal systems and in all of them it plays an important practical role. There are certain legal systems which are considered to be founded on abstract legal principles. Roman law and the legal systems stemming from it, which constitute much of European law, actually belong to this class. In systems such as English law, on the other hand, precedent is clearly the cornerstone of legal thought. In neither system can any new legal term express a definitive meaning until the term itself and its limitations have been tested and evaluated, in other words, until a precedent has been established. To flout a decision taken with regard to an earlier case is to attack the uniformity of the interpretation of legal language and is actually a cause of indecisiveness and probably even of injustice. Every case which is decided should advance the process of defining legal terminology in a way which takes account of precedents and this, of course, should create new precedents. Every term should be tested against local custom and related fields of activity. Judges to whom the duty of interpreting the law is entrusted should carry out their tasks in such a way that if Judge B replaces Judge A there will be no reason to expect a significant change in the court's interpretation of rules and practices. In one sense, this must remain an ideal rather than being an accomplished fact; but unless we follow closely those who uphold these ideals we will have complete disorder; and, what is worse, in a "waste land" unscrupulous persons will try to take advantage of the possibility of interpreting the laws in several ways.

When the proper importance is attached to precedent, if Judge B replaces Judge A the following may be expected:

A. It will be necessary for Judge B to consult Judge A before making a ruling.
B. Judge B, in order to safeguard the law, must make his ruling in accordance with his own views.
C. Only by chance will the two judges concur in any way in their legal interpretations.
D. Judge B's interpretation of the law may be substantially different.
E. Judge B will interpret the law in more or less the same way as Judge A.

Abstract legal principles form the basis of the following:

A. Only English law.
B. Only Roman law.
C. Both English and Roman law.
D. Only the legal systems stemming from Roman law.
E. Roman law and all the systems of law which stemmed from it.

In practice, precedent plays an important role:

A. Mainly in the legal system of the Romans.
B. Only in English law.
C. In the legal systems based on abstract principles.
D. In all legal systems other than the Roman.
E. In all legal systems without exception.

The meaning of a new legal term can become definitive only if:

A. It is in conformity with abstract principles of law.
B. It is carefully examined by a group of legislators.
C. It is used in legal decisions relating to other cases.
D. It is re-interpreted by two or more judges.
E. It is subject to only one legal interpretation.

Precedent is based on the following:

A. Earlier legal decisions.
B. Legal principle.
C. The experiences of judges.
D. Indecisiveness.
E. Roman law.

The meaning of legal terms should reflect not only the law on which they are based but also the following:

A. Abstract principles of law.
B. The knowledge of the judge presiding over the case.
C. Established usage of the time.
D. The uniformity of the interpretation of legal language.
E. The indecisive character of man's affairs.